University of Cincinnati

Date: 3/26/2014

Parasto Alsadat, Kasaie Sharifi, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Business Administration.

It is entitled:
Agent-Based Simulation Modeling and Analysis of Infectious Disease Epidemics and Implications for Policy

Student’s name: Parasto Alsadat, Kasaie Sharifi

This work and its defense approved by:

Committee chair: W. David Kelton, Ph.D.

Committee member: Mark Eckman, M.D.

Committee member: Craig Froehle, Ph.D.

University of Cincinnati
Agent-Based Simulation Modeling and Analysis of Infectious Disease Epidemics and Implications for Policy

A dissertation submitted to the Graduate School of the University of Cincinnati in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in the Department of Operations, Business Analytics, and Information Systems of the Carl H. Lindner College of Business by

Parasto Alsadat Kasaie Sharifi

B.S. Iran University of Science and Engineering March 2014

Committee Chair: W. David Kelton, PhD.
Abstract

Infectious diseases are the largest killers of children and young adults in developing countries, and continue to pose a high burden on global health. This dissertation concerns related topics for combating infectious-disease epidemics, and is organized into three main chapters. While each chapter revolves around a specific topic in the general domain of public-health policymaking, the unifying theme of this research is using agent-based simulation methodology for modeling and analysis of infectious-disease epidemics. The first chapter targets the topic of how best to allocate constrained resources to control an epidemic. In this study, we use a collection of computational techniques, including computer simulation and numerical optimization algorithms, to develop a simulation-optimization framework for addressing the resource-allocation problem as applied to an epidemic. The goal is to relax the restrictive assumptions held by traditional analytical approaches, thereby facilitating a more valid model that is necessarily more complex, and to extend the method to support a general class of resource-allocation problems with realistic assumptions about population structure, disease description, and interventions. The second chapter presents a series of studies on transmission dynamics of tuberculosis (TB), and addresses the impact of various control strategies for combating TB epidemics. Using an agent-based simulation model of a TB epidemic in a household-structured population, we estimate the timing of TB transmission among household and community members. Moreover, we consider multiple case-finding strategies, including household contact tracing and a community active approach, and evaluate the population-level impact of each intervention for controlling disease incidence. Finally, the third chapter analyzes estimation bias of the recent-transmission rate in molecular studies of TB. Analysis of population-based DNA data continues to serve as the main method to estimate the proportion of TB incidence due to recent transmission, which in turn has important implications for understanding the dynamics of transmission and policymaking. Previous studies have identified a number of factors affecting the precision of this approach in various settings, but the exact relationship of such factors remains uncertain. In this study, we aim to quantify the role of such factors, and develop a decision tool for adjusting the estimated ratio of infection due to recent transmission. Using an agent-based simulation model of TB as a virtual laboratory, we implement a sequence of statistically controlled experiments with regard to combinations of several factors. The results enable us to compute the estimation bias for various levels of each factor, and can serve as a decision-support tool for adjusting the estimation error in future molecular studies of TB. In summary, this dissertation concerns critical global-health issues in understanding, controlling, and policy-making concerning infectious-disease epidemics, and offers a multidisciplinary approach to such problems using advanced computer-simulation techniques and analytical tools. The agent-based simulation approach is a novel technique that is increasing in popularity across the literature and in several fields. This brings to bear the power and effectiveness of such models in various applications, and their promising contributions for control and policymaking of infectious diseases.
Acknowledgment

First and foremost, I wish to thank my PhD advisor, Dr. David Kelton, who has been and remains my best role model for a scientist, mentor, and teacher. I have been amazingly fortunate to have an advisor who encouraged me to explore on my own, stood by me in every step of the way, and gave me a hand when my steps faltered. His patience, support and ongoing encouragements helped me to rise above my limits and to discover abilities I never saw in myself. I am grateful to him for holding me to a high research standard, through strict validations of research methods and countless revisions of our manuscripts. Foremost, I thank him for giving me the opportunity of perusing my interest in epidemiological studies despite the conventional research stream at the business program. I will always be in debt to him for supporting me on pursuing my life passion.

My deepest gratitude goes to Dr. David Dowdy, from the Johns Hopkins Bloomberg School of Public Health. I have experienced a limited pivotal moments in my life, and coming in touch with David has been one of them. He introduced me to a new world of research where my breath is taken away with the wonders of science and my eyes shine at the joy of my work. I am deeply grateful to him for our long-distance meetings and discussions that helped me to crank up the problems and to sort out the technical details of my work. His energy and enthusiasm for his research is contagious and motivational, and has helped through tough times in the Ph.D. pursuit. David is more than a mentor, but a dear friend to me, and I cannot be more excited at the sight of my future collaborations with him during the upcoming postdoctoral training.

Moreover, I’m grateful to my dear professor Dr. Marty Levy for his insightful comments and constructive suggestions on statistical aspects of my research. Marty is the funniest
professor in class, and one of the smartest people I know. Our long discussions and his ingenious solutions always leave me with joy and inspiration for learning more from him.

I am also thankful to the members of my dissertation committee, Dr. Craig Froehle and Dr. Mark Eckman, for their insightful comments and feedback on my research, and for their great support and tremendous help with my PhD graduation procedure.

Most importantly, none of this would have been possible without the love and patience of my family back home. They have suffered so much through our separation, but nevertheless, have been a constant source of love, concern, support and strength during all these year. I would like to express my heart-felt gratitude to them: to my mom who raised me with a love of science, and sacrificed so much for supporting me in all my pursuits, to my dad who thought me the first thing about math and computer and for love of whom I never gave up on trying, and finally to my most precious and dearest grandparents whose endless love and devotion gives meaning to my life.

Lastly, I would like to thank my best friend and partner, Parham Mosayebi, for his everlasting support and his unshakeable faith in me. He was the smartest consultant on my problems, the hardest critic of my ideas, and the most patient and loving friend who bore with me through every moment of joy and despair during the past few years. I am truly grateful for his great support and unconditional love.
I dedicate this thesis to

my family for their constant support and unconditional love.

I love you all dearly.
Table of Content

Introduction 12

Chapter 1: A Simulation-Optimization Framework for Allocation of Epidemic Control Resources 19

Essay 1: Toward Optimal Resource-Allocation for Control of Epidemics: An Agent-Based-Simulation Approach 20

1 Introduction 22

2 Problem Statement 23

3 Simulation-Based Approach to Healthcare Resource-Allocation Problems 26
   3.1 Creating an Agent-Based Simulation Model of an Epidemic 26
   3.2 Applying a Simulation-Optimization Technique to Address the Resource-Allocation Problem 28
   3.3 Comparison of the Simulation-based and the Analytical Approaches 30

4 Analysis of a Complex Resource-Allocation Problem 31

5 Conclusion and Future Work 39

6 References 40


1 Introduction 44

2 Background 46
   2.1 RA Model Formulation 48
   2.2 Analysis of the RA Problem 54

3 Methodology 56
   3.1 Step 1: Developing a Simulation Model of the Epidemic. 57
   3.2 Step 2: Incorporate the RA Problem into the ABS Model. 59
   3.3 Step 3: Apply Simulation Optimization to the Freeform RA Problem. 60

4 Illustrative Case 63
   4.1 Simulation Model of Influenza 63
   4.2 The RA Problem in Control of an Influenza pandemic 65
      4.2.1 Calibration, Sensitivity Analysis, and Experiments 65
   4.3 RA Problem Expansion 70
   4.4 Analysis of the Expanded RA Problem 73
      4.4.1 Experiment 1 77
      4.4.2 Further Experiments 81
## Essay 3: Resource Allocation for Controlling Epidemics: Calibrating, Analyzing, and Optimizing an Agent-Based Simulation

1. FluTE Simulation Parameters
2. DALYs Function Calibration ($dq_1$ Sensitivity Analysis)
3. Variance Reduction
   - 3.1 Random-Number Generator (RNG)
   - 3.2 Testing CRN Efficiency
   - 3.3 CRN Experiment
4. Reading the Triangular Coordinates in Mixture Designs
5. Numerical Results
   - 5.1 Experiment 1
   - 5.2 Experiment 2
   - 5.3 Experiment 3
6. References

## Chapter 2: Investigating the Dynamics of Tuberculosis Transmission and Effectiveness of Case-Finding Strategies Using an Agent-Based Simulation

## Essay 1: An Agent-Based Simulation of a Tuberculosis Epidemic: Understanding the Timing of Transmission

1. Introduction
2. Background
3. Method
   - 3.1 Population Structure
   - 3.2 Contact Network
   - 3.3 TB Natural History
   - 3.4 Calibration
4. Results
5. Summary and Discussion
6. References

Supplementary Material
1. Family Probabilities of child bearing 143
2. Contact Network calibration 143
3. Simulation Output in the Base scenario 146

Essay 2: Timing of Tuberculosis Transmission and the Impact of Household Contact Tracing: An Agent-Based Simulation Model 147

1. Introduction 149

2. Methods 150
   2.1 Model and Rationale 150
   2.2 Network and Population Structure 152
   2.3 Natural History of TB 154
   2.4 Calibration 154
   2.5 Interventions and Outcomes 155

3. Results 156
   3.1 Timing of Infections 156
   3.2 Case Finding: Programmatic Indicators 158
   3.3 Case Finding: Impact on TB Incidence 158
   3.4 Sensitivity Analysis 161

4. Discussion 163

5. References 167

Supplementary Material 171

1. Household-Structured Simulation Model of TB 171
   1.1 Population Structure 173
   1.2 Natural History of TB 174
   1.3 Contact Network Formulation 176
      A Note Regarding the Household Transmission Ratio 179
   1.4 Simulation Calibration Procedure 180
      1.4.1 Determining the Initial Population Setting 181
      1.4.2 Computing Simulation Transient Time 181
      1.4.3 Computing the Required Number of Simulation Replications 182

2. Results 183
   2.1 Analysis of the Timing of Infections 185
      2.1.1 Timing of New Infections 185
      2.1.2 Timing of Total Infections 186
      2.1.3 Sensitivity Analysis of Timing of Infection to Variation in the Infectiousness-Versus-Time Relationship 187
      2.1.4 One-Way Sensitivity Analyses: Timing of Infection 188
   2.2 Effectiveness of Case-Finding Interventions 190
   2.3 Sensitivity Analysis of Household Contact Tracing Effectiveness 191
Chapter 3: Estimation Bias of the Ratio of Tuberculosis Incidence Due to Recent Transmission Using DNA Fingerprinting Data

1 Introduction

2 Background
   2.1 Sources of Bias in Molecular Epidemiological Studies of TB
   2.2 Previous Research

3 Mathematical Analysis of Bias in Estimators of the Clustering Level
   3.1 Model Formulation
   3.2 Numerical Analysis
      One-way Sensitivity Analysis of Estimation Bias to the True Clustering Ratio
   3.3 Discussion

4 A simulation Approach for Computing the Estimation Bias of the Recent Transmission Rate
   4.1 An Agent-Based Simulation Model of a TB Epidemic
      4.1.1 Population Structure and Contact Network
      4.1.2 TB Natural History
      4.1.3 Simulating TB Strains
      4.1.4 Simulation Calibration and Verification
   4.2 Simulation Experiments Framework
      4.2.1 Experimental Factors
      4.2.2 Design of Simulation Experiments
      4.2.3 Calibrating Experimental Settings

5 Results
   5.1.1 Effect of Incidence and RCS on Clustering Estimate
   5.1.2 Estimation Bias for Ratio of Incidence due to Recent Transmission
   5.1.3 Reponses-Surface Analysis

6 Conclusion

7 References

Supplementary Material

1. Analytical Model of Clustering Estimation

2. ABS Model Details

3. Calibrating the Experimental Settings
   Choosing the Number of Experiment’s Re-Samples (B)
Introduction

According to the World Health Organization, there is an infectious-disease crisis of global proportions. Infectious diseases are the largest killers of children and young adults, accounting for more than 13 million deaths per year (World Health Organization 1999b). The current threats of pandemic influenza, HIV, and drug-resistant tuberculosis (TB), plus recent threats such as SARS, raise major and urgent concerns with regard to public health preparedness, risk management, and decision-making.

Today’s infectious-disease challenges are broader and more complex than in recent decades (United States General Accounting Office 2001). While the changing, globalized world has brought significant advances for the prevention and control of diseases, it has increased the opportunities for emergence and spread of infectious diseases across the globe. Dramatic increases in the volume and speed of international travel and trade in recent years have increased opportunities for diseases to spread across international boundaries. Diseases once regarded as declining in significance, such as TB, have also reemerged in recent decades to become major global health threats once again. Moreover, the emergence of previously unknown diseases and the development of disease strains resistant to antimicrobial drugs further complicate international disease-control efforts.

The current difficult economic environment, on the other hand, has affected individuals throughout the world. Budgetary and other constraints have in turn had a major impact on public health, requiring difficult decisions at the national, state, and local levels. Ensuring that these important decisions do not affect human health negatively because of weakened public-health capacities will require broad and well-coordinated collaborative efforts to determine the best use of limited resources (CDC 2011).
The ultimate goal of communicable-disease-control programs is to identify risks and prevent excess morbidity and mortality among the population by preventing and managing outbreaks of communicable diseases. Global frameworks for preventing infectious diseases are designed to guide collective public-health action at a time of resource constraints and difficult decisions, while advancing opportunities to improve global health (CDC 2011). Among the critical elements outlined in such frameworks are:

- **Identify and implement high-impact public-health interventions to control infectious diseases:** Focused efforts to prevent and control high-burden infectious diseases can achieve dramatic results within a short time frame. Such efforts can range from identifying and validating new tools for disease prevention and control, to accelerating the uptake and broad use of proven methods for decreasing illness and death from diseases.

- **Develop and advance policies to prevent, detect, and control infectious diseases:** Global protection from infectious diseases requires sound, evidence-based health policies designed to ensure appropriate development and delivery of prevention measures, reduce health disparities, and improve the health of vulnerable populations. Such policies should promote engagement with global partners to reduce cross-border disease spread and contain outbreaks at their source.

To be most effective, these healthcare policies must reflect the best science and public-health thinking, with broad input and consideration of various perspectives, to ensure recognition of the complex societal factors that affect public health. This in turn highlights the important role of interdisciplinary, cross-cutting research into critical health policies and health-care delivery issues, as proposed (and being promoted) by global health organizations such as WHO, CDC, etc. (CDC 2006; World Health Organization 2003)
Epidemiological modeling refers to a family of deterministic/stochastic modeling approaches used for studying the spread of infectious diseases qualitatively and quantitatively. These models have become important tools in analyzing the spread and control of infectious disease, and assist decision makers in taking proper prevention and containment/mitigation measures. The traditional modeling approach, using exact mathematical/analytical methods to describe the course of disease in a deterministic setting, has a long history and includes a large body of research (Anderson & May 1991; Capasso 1993; Kermack & McKendrick 1927). With the advent of better technology and computational power, however, a new branch of computational science, introducing epidemic computer-simulation models, has emerged and has rapidly risen in application. Epidemic simulation models are useful experimental tools for building and testing theories, assessing quantitative conjectures, answering specific questions, determining sensitivities to changes in parameter values, and estimating key parameters from data (Connell et al. 2009; Ma & Xia 2009). Such models are used in assessing the impact of infectious-disease epidemics and pandemics on human health, and their role in comparing, planning, implementing, and evaluating various control programs is of major importance for public-health decision makers.

Various types of computer-simulation techniques have been applied to model and analyze epidemics (Brailsford & Hilton 2001; Rahmandad & Sterman 2008; Connell et al. 2009). A novel simulation approach, agent-based simulation (ABS), is a flexible and powerful approach to studying complex systems involving human interaction. ABS models a system at the individual level by designing and simulating each single person (agent) in the population. The agents are then put into a simulated virtual world representing their living environment (e.g., a neighborhood or city), where they interact with each other and their environment. Introducing a disease pathogen (or carrier vector) to such a model
consequently creates a realistic representation of disease diffusion across time and space. When calibrated to real data from historical epidemics and population-based studies, ABS models can serve as reliable tools for understanding the epidemic dynamics, designing effective containment interventions, and facilitating future research (Burke et al. 2006; Epstein 2009).

In this work, I study agent-based simulation modeling and analysis of infectious-disease dynamics, and its implications for public-health policymaking. My research is organized into several distinct but related studies focusing on different issues in control of infectious disease and public-health policymaking. While each study is developed as a separate piece, the unifying theme of my research is using agent-based simulation methodologies. In this regard, my dissertation is organized into three main chapters based on the relevance and the area of application in each study (Figure 1):

- **Chapter 1**: A Simulation-optimization framework for allocation of epidemic control resources

- **Chapter 2**: Investigating the dynamics of tuberculosis transmission and effectiveness of case-finding strategies using an agent-based simulation approach

- **Chapter 3**: Estimation bias of the ratio of tuberculosis incidence due to recent transmission using DNA fingerprinting data
The first chapter revolves around the topic of how best to allocate constrained resources to control an epidemic. This includes my earlier research during the first two years of the PhD program on developing a simulation-optimization framework for a general instance of epidemic resource allocation problem. In this study, we use a collection of computational techniques, including computer simulation and numerical optimization algorithms, to develop a simulation-optimization framework for addressing the resource-allocation problem as applied to an epidemic. The goal is to relax the restrictive assumptions held by traditional analytical approaches, thereby facilitating a more valid model that is necessarily more complex, and to extend the method to support a general class of resource-allocation problems with realistic assumptions about population structure, disease description, and interventions. The results of this study have been presented and published at several meetings (Kasaie et al. 2010), and have been subsequently published through two refereed-
journal papers in *IIE Transactions on Healthcare Systems Engineering* (Kasaie & Kelton 2013a; Kasaie & Kelton 2013b).

In the second chapter, I present a series of studies on transmission dynamics of tuberculosis, and study the impact of various control strategies for combating TB epidemics. This research was conducted as a joint research activity during my parallel studies in the Business Administration doctoral and Biostatistics master’s programs from 2011 to 2014. In this work, we initially propose an agent-based simulation model of TB for studying the dynamics of transmission across various social networks, and estimate the timings of secondary infections. Based on our findings, we then refine our modeling assumptions and propose a new simulation model of a TB epidemic in a household-structured population, which is subsequently used to estimate the timing of TB transmission among household and community members, and to evaluate the likely impact of various case-finding strategies for controlling TB. The preliminarily section of this work has been presented as a part of my MS thesis in Biostatistics (Kasaie 2014), and the results were presented and published (refereed) in the 2013 Winter Simulation Conference (Kasaie et al. 2013). Moreover, an original research manuscript based on the later phases of this study has recently received final acceptance for publication at the *American Journal of Respiratory and Critical Care Medicine* and is currently in press (Kasaie et al. 2014).

The third chapter of my dissertation reports the results of recent research in the field of TB molecular biology, and targets the estimation bias of the TB recent-transmission rate using DNA fingerprinting data. Cluster analysis of population-based DNA data continues to serve as the main method to estimate the proportion of TB incidence due to recent transmission, and this ratio in turn has important implications for understanding the dynamics of transmission and policymaking. Such an estimation procedure, however, is prone to many factors affecting the precision of final results, and previous studies have
identified a number of such factors in separate settings. In this study, we aim to quantify the role of such factors in the presence of each other, and measure the estimation bias (from its true value) in a variety of settings. Our results can ultimately be used as a decision-support tool for adjusting the estimated ratio of recent transmission in future epidemiological studies of TB (e.g., researchers can plug in their various factors describing study specifications and epidemic’s characteristics, and compute the expected value of bias for their inference).

Using an agent-based simulation model of TB as virtual laboratory, we implement a series of statistically controlled experiments with regard to combination of several factors. The results enable us to compute the estimation bias for various levels of each factor, which can serve as a decision-support tool for adjusting the estimation error in a variety of settings.

In the following, I present the three research chapters, followed by the concluding discussion in Chapter 4. The first two these ensuing chapters are organized into a collection of separate essays, each presenting a published or accepted manuscript, and the last chapter is presented as a single research manuscript that has not yet been submitted for publication. Literature reviews are presented separately for each essay, and each essay is followed by a corresponding list of references and supplementary material.
Chapter 1: A Simulation-Optimization Framework for Allocation of Epidemic Control Resources
Essay 1: Toward Optimal Resource-Allocation for Control of Epidemics: An Agent-Based-Simulation Approach

Parastu Kasaie¹, W.David Kelton¹, Abolfazl Vaghefi², S.G.R. Jalali Naieni³

¹Department of Quantitative Analysis and Operations Management, University of Cincinnati, Cincinnati, Ohio, 45221, U.S.A
²Department of Industrial and System Engineering, Rutgers University, Piscataway, NJ, USA
³Department of Industrial Engineering, Iran University of Science and Technology, Tehran, 16844, Iran

Abstract

Employing mathematical modeling and analytical optimization techniques, traditional approaches to the resource-allocation (RA) problem for control of epidemics often suffer from unrealistic assumptions, such as linear scaling of costs and benefits, independence of populations, and positing that the epidemic is static over time. Analytical solutions to more realistic models, on the other hand, are often difficult or impossible to derive even for simple cases, which restricts application of such models. We develop an agent-based simulation model of epidemics, and apply response-surface methodology to seek an optimum for the RA output in an iterative procedure. Validation is demonstrated through comparison of the results with the mathematical solution in an RA example for which the analytical solution is known. We apply the proposed approach to a more complicated RA problem in which a number of previous restricting assumptions are relaxed.
1 Introduction

Epidemics of infectious diseases such as influenza, malaria, and human immunodeficiency virus (HIV) are a major threat for social health. While the World Health Organization (WHO) reports expenditure of more than $500 million in 2008-2009 to control the spread of epidemics (World Health Organization programme budget 2009), the demand for efficient allocation of these resources is growing. The resource-allocation (RA) problem concerns the best strategy for policy makers to allocate a fixed budget to various populations, through targeted interventions that affect the epidemic’s parameters. The efficacy of an intervention on the value of epidemic parameters is usually an increasing function of the cost of intervention and referred as the cost function (Brandeau et al.2003).

Allocation of epidemic-control resources has been studied for many years, and the literature contains analytical models using linear, integer, and dynamic programming (Epstein et al. 2005; Van Zon and Kommer 1999; Earnshaw and Dennett 2003). These models, however, are not applicable to epidemics with non-linear rates of growth and are restricted by several other assumptions like the number of interventions or independence of populations. Recent mathematical approaches to healthcare-resource allocation, on the other hand, suggest advanced models of disease prevalence among several populations, and consider more general forms of a cost function for prevention programs. Zaric and Brandeau (2001) suggest heuristic algorithms for solving RA problems, and approximating the epidemiological system when closed-form solutions are not known. However, further augmentation in the scope of the problem for real cases could rapidly increase the number and complexity of equations, so the final models might become intractable even for simple instances.
Agent-based simulation (ABS) models are powerful tools that can describe structured epidemiological processes involving human behavior and local interactions. While the computational capacity of ABS models allows for developing large-scale models of epidemics, they are flexible enough to display detailed and complex characteristics of a real system. ABS models have been used to simulate epidemics and assess policy options (Longini et al. 2004; Epstein et al. 2007). Such models represent the system behavior at both macro and micro levels, and allow investigation into system behavior, sensitivity analysis, and predictions.

In this paper, we present an ABS-based approach for allocation of epidemic-control resources. Developing an ABS model of epidemics, we investigate the response surfaces of RA objectives, and apply statistical simulation-optimization techniques to search for the optimal allocation of available resources.

We introduce the RA problem in Section 2. In Section 3 we present the ABS approach, discuss the application of simulation-optimization techniques to address the RA problem, and compare the ABS approach with the analytical approach for a case when the latter is workable, to validate that the ABS approach agrees. In Section 4, we illustrate the efficiency and applicability of the ABS approach thorough an RA example that is much more complex and cannot be solved by analytical methods. The conclusion is provided in Section 5.

2 Problem Statement

Consider an epidemic of a single disease existing in $p$ populations (e.g. countries, cities, or any other groups of people under study). In order to model the progress of an epidemic within each population, the diversity of individuals in various fields must be reduced to a few key characteristics. This is done by dividing each population into subgroups, also called compartments. Each compartment consists of individuals in a specific disease state (e.g.
susceptible, infected). Transmission of the disease may occur through one or more diverse pathways, but in this paper we consider only transmission through physical contact with infected individuals. Additional assumptions regarding the epidemic process differ by study.

Disease outbreak is usually far more rapid than the natural vital dynamics of the population (natural births and deaths, migrations, etc.), so that one may neglect them. In this case disease prevalence can be modeled through a set of ordinary differential equations, initially proposed by Anderson and May (1991). In general, however, the timeline of the study may extend to several years due to the nature of the disease, or the horizon of policy making. The epidemic process, in this case, is composed of several aspects of population dynamics, and the associated models incorporate additional epidemic parameters, such as the rate of natural birth/death, rate of migration into and out of populations, etc. These parameters are usually determined by characteristics of the disease and the population under study, and can be defined in stochastic form, or as a function of other parameters.

Resources used for combating a disease are assumed to affect parameters of the epidemic model (e.g. a specific therapy can reduce the disease progression rate) through healthcare interventions. Healthcare interventions target epidemic parameters in a specific compartment or in an entire population. Associated with each intervention is a cost function that depicts the relationship between the amounts invested in the intervention and the values of the associated parameters in the epidemic model. For a total available budget of $B$, let $v_h$ be the amount of money invested in intervention $h$, $h = 1, 2, \ldots, n$, where $0 \leq v_h \leq B$, and let $v = (v_1, v_2, \ldots, v_n)$ be the investment vector. We define $H(v)$ as the objective function of the RA problem, with investment values of $v_h$ as the decision variables. The general form of the RA problem is
\[ RA: \text{Optimize} H(v) \]
\[ S.t: \sum_{h=1}^{n} v_h = B \]
\[ v_h \geq 0 \quad \forall h \]

The epidemiology literature contains several discussions on the appropriate form of objective functions in health-care policy making (Phillips et al. 1998). Although the definition and types of objective functions are based on several factors (such as the disease under study, characteristics of the population, and scope of decision making), there are general guidelines in choosing the appropriate function. However, analytical approaches to epidemiological problems are often restricted in form and the number of objective functions, and may become too complex or even intractable for nonlinear or dynamic cases. In a simulation-based approach, on the other hand, models of epidemics provide a virtual reality to generate any desired outputs, and simulation-optimization techniques put no restriction on the form or nature of objective functions. We consider two generally accepted forms of objective functions suggested by Brandeau et al. (2003). The first is to minimize the number of new infections occurring during the time of the study, \( INF(v) \), and the second is to maximize the total number of quality-adjusted life years, \( QALY(v) \), gained. Let \( q_{ij} \in (0, 1), i = 1, 2, \ldots, p; j = 1, 2, \ldots, m \) denote the quality adjustment for life years lived by individuals in compartment \( j \) of population \( i \). We assume that quality of life is higher for individuals in the earlier state of disease than for individuals in late states; thus \( q_{ij} > q_{ij'} \) for \( j' > j \). For more information on these functions, see Brandeau et al. (2003).
3 Simulation-Based Approach to Healthcare Resource-Allocation Problems

In this section, we discuss our simulation-based approach to address the healthcare resource-allocation problem. This approach consists of two major steps: creating the ABS model of an epidemic, and applying a simulation-optimization technique to estimate the RA problem. We close this section by investigating the consistency of results with the analytical solution for a relatively simple example where the analytical solution is available, by way of validation of our approach.

3.1 Creating an Agent-Based Simulation Model of an Epidemic

Agent-based modeling and simulation is a relatively new modeling paradigm that has seen extensive application in recent years. While discrete-event simulation (DES) is still more common in operations research, ABS introduces a new way to model complex systems. Such systems are characterized by the fact that their aggregate properties cannot be deduced simply by looking at how each component behaves, since the interaction structure itself is playing a crucial role. In comparison with the top-down modeling approach of DES (where a system is broken into its components represented by blocks, machines, or modules, and entities are defined as passive objects being directed through these components), ABS follows a bottom-up approach. In ABS, a system is modeled as a collection of autonomous decision-making entities called agents. Each agent individually assesses its situation, and makes decisions on the basis of a set of rules. Agents interact with one another, and with the environment through a computer code. Over many replications, these interactions can generate large-scale phenomena of interest, in our case the course of epidemics across space and time. This generative nature of such models enables us to focus on the microscopic individual behavior, as well as study the macroscopic pattern of epidemics emerging in a
larger scale. In this regard, as Burke et al. (2006) suggest, ABS can provide credible bases for policy analysis when calibrated to actual epidemic data.

We choose NetLogo (2010), a popular agent-based programming language that is particularly designed for modeling complex systems developing over time. Figure 1 shows the proposed ABS logic of an epidemic model implemented as a set of five main sub-procedures: Creation, Contact, Progression, Migration, and Reproduction.

The simulation model is initialized by defining the model global variables and each agent’s attributes. In the create procedure, the epidemic system is created as a collection of agents in \( p \) different populations. Each population consists of \( m \) compartments with agents in different states of disease, where the initial size (number of agents) of the \( j^{th} \) compartment of the \( i^{th} \) population is \( N_{0ij}, i = 1, 2, \ldots, p; j = 1, 2, \ldots, m \). The contact procedure simulates the process of disease prevalence in each population. Infected individuals in later states of disease can transmit the disease to susceptible individuals (in the first compartment) through random contacts where \( \lambda_{ij} \) is the sufficient contact rate for transmission among individuals of the first and the \( j^{th} \) compartment of population \( i \). With progression of the disease at the next procedure, the individuals in the \( j^{th} \) state of disease move to the next state with probability \( \theta_{ij} \). Migration takes place among individuals in different populations but in the same disease state \( j \), with probability \( \varphi_{ii'j}, i \neq i' \). In the Reproduction procedure, a number of current agents in different populations and disease states die with probability \( \delta_{ij} \) for each agent. The remaining agents in each population then will have probability of \( \zeta_{i} \) to bear new children, who will belong to their parent’s population, but do not inherit the disease. The model is executed until the simulation time reaches the study time horizon of \( T \). The model outputs are defined as \( \text{INF}(v) \) and \( \text{QALY}(v) \) for individuals in each population. The outputs are reported at the end of each replication.
3.2 Applying a Simulation-Optimization Technique to Address the Resource-Allocation Problem

By simulation-optimization we mean a repeated analysis of the simulation model with different values of input parameters, in an attempt to identify the best simulated system performance (Barton and Meckesheimer 2006). However, for the extensive experimentation required for optimization, the simulation models themselves may require excessive computation, so simpler approximations are constructed, often referred to as meta-models (Kleijnen 2008) or surrogate models (Yesilyurt and Patera 1995). A meta-model, or a model of model, provides a concise representation of the output response and its dependence on accompanying input factors. A meta-model simplifies the simulation-optimization in two ways: the meta-model responses are deterministic rather than stochastic, and the run times are generally much shorter than the original simulation. However, the meta-model is not an exact replica of the simulation model, so there is a trade-off involved.
Meta-model-based optimization methods use an indirect-gradient optimization strategy to seek the optimal solution. *Response-surface methodology* (RSM) is a collection of mathematical and statistical techniques that are useful for the modeling and analysis of problems in which a response of interest is influenced by several variables and the objective is to optimize this response. RSM is a meta-model-based optimization heuristic that fits first- or second-order polynomial regression models to observed values of $Y$, the simulation output. An example of a full second-order response surface model would be

$$Y(\theta) = \beta_0 + \sum_{j=1}^{q} \beta_j \theta_j + 2 \sum_{i=1}^{q} \sum_{j=1}^{i-1} \beta_{ij} \theta_i \theta_j + \sum_{i=1}^{q} \beta_i \theta_i^2 + \epsilon,$$

where $\epsilon$ is an independent normal random variable with mean 0 and variance $\sigma^2$. Initiating from a randomly selected or predetermined local region, RSM designs the appropriate simulation experiment, typically factorial designs for first order, and central composite designs (CCDs) for second order, and fits a local meta-model to the response. This model is then used to decide the direction of improvement called *steepest ascent* (or *descent*). Investigations continue along the steepest direction until no further improvement in simulation output is observed. The procedure then moves to the next iteration by replacing old meta-models with new ones, and improving the result following the steepest direction. The stopping criteria are checked at the end of each search iteration, and the optimal solution is estimated at the end of final iteration. We consider failure to achieve a minimum of 5% improvement in average response as the stopping criterion of our procedure (Castillo 2007).
The RA problem seeks the best strategy to invest a fixed budget among populations through targeted interventions, with the goal of optimizing the problem objective functions. In our model, the simulation outputs, $\text{INF}(v)$ and $\text{QALY}(v)$, represent the objective functions controlled by the investment vector of $v = (v_1, v_2, \ldots, v_n)$. As defined in Section 2, $v_h, h=1, 2, \ldots, n$, is the invested amount of money in intervention $i$, which is designed as the input of the simulation model. In order to solve the RA problem, and to estimate how to optimize the simulation’s outputs, we apply RSM to our ABS model in an iterative procedure.

### 3.3 Comparison of the Simulation-based and the Analytical Approaches

Brandeau et al. (2003) formulated the problem of resource allocation among non-interacting populations in general, and established conditions that characterized the optimal solution in certain cases. We apply our approach to a numerical example of such a problem, and compare our solution with the results of the exact RA mathematical model, which demonstrates consistency of our simulation-based results and analytical solution.

Assume an epidemic among four non-interacting populations ($p = 4$), with constant sizes of $N_i; i = 1, 2, \ldots, 4$ over time. The epidemic within each population is described by a basic susceptible/infected (SI) epidemic model with $I_0i$ and $S_0i$ denoting the initial proportion of susceptible and infected individuals in population $i$. The natural rate of birth and death, $\Delta_i$, is the same for both infected and susceptible individuals in each population. The total amount of available funds is $B$, which can be spent to affect the contact rate among individuals. Therefore, the cost function $c_i(\lambda_i)$ denotes the net present cost of immediately achieving a sufficient contact rate $\lambda_i$ in population $i$. The cost functions are assumed to follow non-linear growth over time, independent from each other, and to be strictly decreasing in $\lambda_i$, and $c_i(\lambda_{0i})$
= 0, \( i = 1, 2, \ldots, 4 \); where \( \lambda_{0i} \) is the initial contact rate of individuals in population \( i \) at time zero.

Considering the objective function of minimizing the number of new infections, \( \text{INF}(v) \), we assumed that all epidemic parameter values are continuously uniformly distributed, and applied the simulation-based approach to the RA problem, as discussed above. Subsequently the mathematical model was solved using LINGO (2010), which showed the same results with the solution given by simulation optimization. The results verify the performance of our ABS-RSM approach in this case to represent the epidemic system well, and demonstrate the consistency of our proposed approach against the analytical solution of at least this well-known RA problem. However, the precision of the simulation-optimization’s results may still vary due to the random nature of simulation runs, scale of the model, or complex behavior of the response surface in more complicated types of RA problems. This may consequently require further analysis of the response surface to estimate the optimal solution, and statistical hypotheses to test the optimality of the suggested solution. In the following section, we demonstrate the applicability and efficiency of the ABS approach in a more complicated RA problem for which an exact analytical optimum will be extremely hard to derive.

4 Analysis of a Complex Resource-Allocation Problem

In this section, we apply our approach to a more complicated RA problem in which a number of previous restricting assumptions (e.g. independence of populations\ interventions, constant value of epidemics’ parameters over time, equal rates of birth and death, etc.) are relaxed. This example demonstrates how the proposed method can effectively be used in more realistic epidemic models and complex RA problems for which deriving the analytical
solution may be impossible. This example was designed with regard to a similar model proposed by Zaric et al. (2001).

Consider an epidemic among $p=2$ populations as in Figure 2. The fixed budget is invested in three different interventions affecting the epidemic parameters. The RA objective is defined as maximizing $QALY(v)$ while maintaining an upper bound for the value of $INF(v)$ at the end of the time horizon. This requirement for the value of $INF(v)$ can eventually barricade the RSM at the boundaries of the feasible region. Luckily for us, such a problem didn’t occur.

![Figure 2: A Three-State Epidemic Model Among Two Interacting Populations](image)

In this example, populations represent high-risk and low-risk groups of people in a society (e.g. the first population may represent intravenous drug users with a higher risk of disease transmission), and migration can take place among individuals in the same compartments of different populations. The epidemiological system consists of three states, i.e. susceptible,
early infected, and late infected, with different epidemic parameters. It is also assumed that only the infected individuals in the late state can transmit the disease to uninfected individuals.

Three types of healthcare interventions are designed to control the spread of the disease. Each intervention targets one of the epidemic parameters $g_h$, $h = 1, 2, 3$, which are assumed to be the migration rate from a high-risk to low-risk group ($\phi_{12}$), and the individual contact rates ($\lambda_{12}, \lambda_{22}$) in each population. Associated with each intervention is a cost function

$$W_{g_h}(v) = M_h(v) \cdot W_{g_h}(0) \quad h = 1, 2, 3$$

where $W_{g_h}(v)$ is the future value of epidemic parameter $g_h$ when investing the vector $v = (v_1, v_2, v_3)$ among interventions, $W_{g_h}(0)$ is the initial value of this parameter, and $M_h(v)$ is a nonlinear function of the form

$$M_h(v) = \alpha_h + \beta_h \exp(-\gamma_h \times v_h) + \eta_h \exp(-\gamma_{h+1} \left(\frac{v_{h+1}}{a_h}\right)^{b_h})^{-1} \quad h = 1, 2, 3.$$ 

The first part of this function, $P1$, models the nonlinear effectiveness of each intervention, where $\alpha_h$ is the location parameter taken from a continuous $[0, 1]$ uniform distribution, $\beta_h$ is a shape parameter with similar values among interventions, and $\gamma_h$ is a coefficient weighting the amount of investment in each prevention program. Figure 3 demonstrates the nonlinear trend of $P1$ in each intervention for different amounts of investment.
Interventions can be thought of as risk-reduction programs in each population. However, the populations are not independent from each other, and the effects of an intervention are not necessarily restricted to the target group. In other words, interactions may occur among interventions, so that the amount invested in one prevention program could influence the effectiveness of another program. For example, consider an epidemic of a viral disease with a higher risk of infection among smokers. A public prevention program is designed to control the rate of disease transmission among the low-risk population. Such a program not only increases the social knowledge about the disease nature and reduces the transmission rate among the low-risk group, but also influences the social norm toward risky behaviors (smoking). This can consequently affect the individuals in the high-risk population to reduce their risky behaviors (quit smoking) and increase the rate of migration from the high-risk group to the low-risk group. This interaction is modeled through the second part of $M_h(v)$ ($P_2$) with a nonlinear return to scale. We assume that investment in intervention $h$ ($h = 2$ or $3$) can influence the effectiveness of intervention $h - 1$, and consequently improve the value of the parameter $g_{h-1}$. In this part, the coefficient $\eta_h$ is taken from a continuous $[0, 1]$ uniform distribution, and the values of $a_h$ and $b_h$ are used to scale the strength of the
interaction. The associated values of these cost functions’ parameters for each intervention are in Table 1.

<table>
<thead>
<tr>
<th>h</th>
<th>Cost function</th>
<th>$W_0(0)$</th>
<th>$a_h$</th>
<th>$\beta_h$</th>
<th>$\gamma_h$</th>
<th>$\eta_h$</th>
<th>$a_h$</th>
<th>$b_h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$C(\varphi_{121})$</td>
<td>0.112</td>
<td>0.149</td>
<td>0.701</td>
<td>-0.0002128</td>
<td>0.15</td>
<td>100</td>
<td>1.85</td>
</tr>
<tr>
<td>2</td>
<td>$C(\lambda_2)$</td>
<td>0.5</td>
<td>0.414</td>
<td>0.786</td>
<td>0.00042126</td>
<td>-0.2</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>$C(\lambda_1)$</td>
<td>0.7</td>
<td>0.293</td>
<td>0.707</td>
<td>0.00046151</td>
<td>0</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

We assume a special type of disease that could reduce the pregnancy chance of infected individuals, and assume that the probability of child bearing is an exponential function of disease duration. Moreover, the disease progression rate of $\theta_{ij}, i = 1, 2; j = 1, 2, 3,$ for each individual is assumed to be a function of disease duration (see Table 2 for formulae and values of parameters), where $k$ is a constant coefficient associated with the severity of disease in each population (early progression of disease is faster in the high-risk population). We also assume exponential death rates of $\delta_{ij}, i = 1, 2; j = 1, 2, 3,$ with different values for individuals in each compartment (Bailey 1975), and define other epidemic parameters as shown in Table 2.

35
### Table 2: Notation and Parameter Values for RA example

<table>
<thead>
<tr>
<th>Indices</th>
<th>Description</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i, i'$</td>
<td>Indices for population $i = 1, 2$</td>
<td></td>
</tr>
<tr>
<td>$j, j'$</td>
<td>Indices for epidemic model compartments, $j = 1, 2, 3$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Global Parameters</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$B$</td>
<td>Total budget</td>
<td>10000</td>
</tr>
<tr>
<td>$T$</td>
<td>Time horizon</td>
<td>20</td>
</tr>
<tr>
<td>$N_{1j}$</td>
<td>Size of compartment $j$ of population 1</td>
<td>500, 300, 200</td>
</tr>
<tr>
<td>$N_{2j}$</td>
<td>Size of compartment $j$ of population 2</td>
<td>1000, 600, 400</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Epidemic Parameters</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varsigma_i$</td>
<td>Entrance rate of population $i$</td>
<td>$0.25 \exp(-\frac{t_{\text{infection}}}{10})$</td>
</tr>
<tr>
<td>$\theta_{ij}$</td>
<td>Disease progression rate in population $i$</td>
<td>$1-\exp(-\frac{t_{\text{infection}}}{k_j})$;</td>
</tr>
<tr>
<td>$k_i$</td>
<td>Constant coefficient of disease progression</td>
<td>$k_1 = 4, k_2 = 20$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\exp(d_j); d_j = 0.2, 0.23$,</td>
</tr>
<tr>
<td>$\delta_{ij}$</td>
<td>Death rates in compartment $j$ of population 1</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\exp(d_j); d_j = 0.18, 0.21$,</td>
</tr>
<tr>
<td>$\delta_{2j}$</td>
<td>Death rates in compartment $j$ of population 2</td>
<td>0.24</td>
</tr>
<tr>
<td>$\varphi_{21}$</td>
<td>Migration rates from population 2 to 1 for in the first compartment</td>
<td>0.4</td>
</tr>
<tr>
<td>$q_{1j}$</td>
<td>Quality adjustment for life years in compartment $j$ of population 1</td>
<td>0.32, 0.17, 0.1</td>
</tr>
<tr>
<td>$q_{2j}$</td>
<td>Quality adjustment for life years in compartment $j$ of population 2</td>
<td>0.55, 0.25, 0.15</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simulation Model Variables</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>Time of simulation (simulation clock)</td>
<td></td>
</tr>
<tr>
<td>$t_{\text{infection}}$</td>
<td>The time of infection</td>
<td></td>
</tr>
<tr>
<td>$deads$</td>
<td>Total number of deaths</td>
<td></td>
</tr>
</tbody>
</table>

We let $v = (v_1, v_2, v_3)$ be the investment vector, and develop an ABS model of epidemics with the $v$ as the input and the QALY($v$) and INF($v$) as the outputs reported at the end of each simulation run. The final goal is to determine the inputs that maximize the total value of QALY($v$) at the end of time horizon, while maintaining the upper bound of 12500 for the value of INF($v$). Considering the binding constraint of the total budget, the RA problem is
\( RA: \text{Maximize } QALY(v) \)

\[
\begin{align*}
S.t: \quad & \sum_{h=1}^{3} v_h = 10000 \\
& v_h \geq 0 \\
\text{Requirement } \inf(v) \leq 12500
\end{align*}
\]

The RA problem analysis begins with applying RSM to the simulation model. We start our investigation with the initial vector of \( v = (2000, 5000, 3000) \) for the invested amounts in each intervention, and proceed with the investigation in an iterative procedure. A \( 2^k \) factorial design is used at the first iteration to obtain a linear meta-model of outputs. The design is augmented with five center points that allow checking the adequacy of the fitted polynomial. The number of simulation replications was 300 during the first iteration; however, this number went up to 1700 replications for the final experiments to obtain a relative precision of 5% for a 95% half-width interval over the point estimation. Investigations were conducted along the direction of maximum improvement in response until no further increase in \( QALY(v) \) point estimation was observed. Moreover, the value of \( \inf(v) \) is checked at each step to assure the requirement of meeting the upper bound. At this stage, the performance of the suggested optimum is checked, and the searching continues in a new iteration if needed. The RSM results are provided in Table 3.
The final approximate optimum is identified through the 6th RSM iteration for an investment vector of \( v = (5673.8, 826.2, 3500) \). A CCD experimental design with 5 center points is used to check the performance of the simulation model at this point. We also check the value of INF\( (v) \) as the second priority of optimization. Figure 4 shows the overlay contour plot of both responses (QALY\( (v) > 16800 \) and INF\( (v) < 12400 \)) for this experiment. The black dot in this figure demonstrates the approximate stationary point of \( v = (5708.8, 802, 3491) \) with the corresponding value of QALY\( (v) = 16800.2 \) and INF\( (v) = 12188.1 \) for the outputs.
5 Conclusion and Future Work

This paper presented an agent-based-simulation approach for allocation of epidemic-control resources. The proposed approach considers diverse resource-allocation problems with only general and weak assumptions made about the shape of the cost function and the underlying epidemic structures. Applying optimization-approximation techniques to the ABS model of epidemics, we solved the RA problem in a stepwise procedure. We demonstrated the consistency of our results with an analytical solution through a simplified RA example for which analytical results were previously derived. The application of the suggested approach is finally discussed in a more complex and realistic RA example for which deriving an analytical solution might be impossible.

Use of the ABS approach introduces several advantages to this type of research. Compared to other more-common simulation approaches such as discrete-event simulation where modeling is done at the macroscopic level and entities are just passive objects flowing through block diagrams of the model, ABS allows us to design detailed individual behaviors and their interactions at the microscopic level, so that the developed models will eventually provide a valid representation of population dynamics and disease prevalence through the course of time. The flexibility of the developed model, on the other hand, enables us easily to incorporate new assumptions about populations’ characteristics and disease characteristics. We developed our ABS models using the NETLOGO software, which despite the ease of programming, suffers from a lack of statistical or optimization tools’ support for analysis of the simulation output.

Future work includes migrating the simulation model of epidemic to other commercial ABS platforms such as REPAST, development of a comprehensive simulation model of epidemics with other means of transmission, extension of our optimization approach to
dynamic RA problems, and more robust statistical testing of optimality conditions for the derived solution.

6 References


NetLogo homepage available via <http://ccl.northwestern.edu/netlogo/> [accessed March, 1, 2010].


Parastu Kasaie, W. David Kelton

Department of Operations, Business Analytics, and Information Systems, University of Cincinnati, PO Box 210130, Cincinnati, Ohio 45221-0130, USA

Published in IIE Transactions on Healthcare Systems Engineering 3, no. 2 (2013): 78-93
Abstract

We consider the problem of resource allocation (RA) in the control of epidemics where a fixed budget is allocated among competing healthcare interventions to achieve the best health benefits, and propose a simulation-optimization framework to address a general form of the problem. While traditional approaches to the epidemic RA problem suffer from restrictive assumptions to facilitate exact analytical solutions, a simulation-based technique relaxes such assumptions and provides a more realistic representation of the epidemic. Coupling the simulation model with optimization techniques enables us to analyze the behavior of RA outcomes with regard to different investment strategies, and seek optimal allocations. We discuss implementation steps and illustrate our approach for an RA problem in the control of influenza pandemic with several interacting healthcare interventions.

Key words: Epidemic modeling, Resource allocation, Simulation optimization, Response-surface methodology
1 Introduction

According to the World Health Organization, there is an infectious disease crisis of global proportions. Infectious diseases are the largest killers of children and young adults, accounting for more than 13 million deaths per year (World Health Organization 1999a). Although infectious diseases pose a serious threat to public health, resources for their control are limited. Decision makers must determine how to allocate limited epidemic-control budgets among competing healthcare interventions to optimize health benefits. Healthcare interventions include vaccination, prevention, and treatment programs, and target epidemic parameters in a specific disease compartment, or in an entire population. The effectiveness of an intervention is usually measured by changes in an epidemic parameter’s value rather than the outcome, and is modeled as an efficacy function.

Resource allocation (RA) for epidemic control is far more complex than typical RA problems, and traditional approaches are typically limited in their ability to address such complexities. Traditional approaches use a mathematical formulation of the epidemic model and apply analytical techniques, such as linear/integer programming, optimal control, or equilibrium analysis. These approaches, however, suffer from the simplifying assumptions they make regarding the populations’ structure, the form of efficacy functions, and the model’s parameters. Recent analytical approaches consider more general instances of RA problems and relax some assumptions. However, the models are often too complex to remain tractable for large, realistic cases. Researchers have analyzed specific instances of such models, and proposed optimization heuristics to approximate the solution in more general cases (see Section 2.2).
Various types of computer simulation have been applied to epidemics. Computer simulation and in particular *agent-based simulation* (ABS) is a flexible and powerful tool in modeling complex systems involving human interaction. Researchers have developed ABS models of specific diseases that have been used for policy making, but they are often restricted to just sensitivity analysis, as opposed to optimization, with respect to the model’s parameters, or comparison of a small set of fixed, given alternative strategies.

We too follow a simulation approach but propose a simulation-optimization framework. We extend our previous work in (Kasaie et al. 2010), and discuss practical aspects of our methods and implications of our solutions to challenges arising in practice. Our approach relaxes the restrictive assumptions of analytical approaches, and supports general models of epidemics with realistic population structure and disease history. The simulation-optimization technique poses no particular restriction on the form or shape of efficacy functions, and enables analysis of complicated instances of the RA problem.

Providing a general background, we formulate our RA problem in Section 2 and discuss the challenges in its analysis. In Section 3, we present our methods and illustrate implementation steps. We provide a typical example of a hypothetical RA problem in the context of an influenza pandemic, and analyze an instance of this problem using our suggested approach in Section 4. This illustrates application of our work to a complicated problem with multiple interacting interventions, multiple objective functions, and a realistic epidemic model structure. We discuss some practical insights and conclude in Section 5. Additional details on the calibration, parameterization, and optimization analysis of our ABS model are provided in the companion paper (Kasaie & Kelton 2013a).
2 Background

The dynamics of infectious diseases have an old and rich history in mathematical biology. Epidemiological modeling refers to a family of modeling approaches used for studying the spread of infectious diseases qualitatively and quantitatively. Compartmental modeling is one of the popular dynamic modeling approaches mostly suited for study of microparasitic infections. In compartmental modeling, a population is divided into disjoint classes, corresponding to different disease states, whose sizes change with time $t$. The total population size $n_{\text{total}}$ can be constant or variable, but assumed large enough so that the compartment sizes can be regarded to change continuously over time.

Compartmental-modeling approaches represent the epidemic using a system of equations. For a simple example of such models, consider a deterministic SIR compartmental model with three compartments: susceptible ($S$), infected ($I$), and recovered ($R$) individuals. Under a homogeneity assumption, the rate of new infections is $\lambda(t)S(t)I(t)$ where at time $t$, $S(t)$ is the number of susceptible individuals, $I(t)$ is the number of infected individuals, and $\lambda(t)$ is the infection coefficient, e.g., a sufficient contact rate for transmission of the disease at time $t$. Similarly, the recovery rate at time $t+1$ is $\gamma(t)I(t)$ where $\gamma(t)$ is the recovery coefficient. Then the model is:

\[
\begin{align*}
\frac{dS(t)}{dt} &= -\lambda(t)I(t)S(t), \\
\frac{dI(t)}{dt} &= \lambda(t)I(t)S(t) - \gamma(t)I(t), \\
\frac{dR(t)}{dt} &= \gamma(t)I(t)
\end{align*}
\]

where $S(0) = S_0 \geq 0$, $I(0) = I_0 \geq 0$, $R(0) = R_0 \geq 0$.

Adding the assumptions that the population’s size is fixed, i.e., $S(t) + I(t) + R(t) = n_{\text{total}}$ where $n_{\text{total}}$ is fixed, and constant values of epidemic parameters over time, i.e., $\lambda(t) = \dot{\lambda}$ and
\( \gamma(t) = \gamma \), simplifies the model to a system of ordinary differential equations, which can be solved numerically by performing the computation at various time-step lengths and then picking a time step short enough to reduce the computational error to be below a selected threshold \( t \) (Anderson & May 1991).

The formulation of compartmental models depends on the natural history of the disease and general characteristics of the population. The spread of an infectious disease can involve further disease-related factors (e.g., the mode of transmission, the latent and infectious periods, or susceptibility and resistance), as well as additional assumptions regarding heterosexual transmission, host-vector groups, multiple mixing groups, population age structure, etc. To include such assumptions, the basic structure of our model can be extended to more complex stochastic compartmental models, addressing continuous-time Markov, and non-Markov processes, and to include several sub-compartments presenting the population heterogeneities. Inclusion of additional assumptions, however, is not without cost, as the resulting models are often too complex and analytically intractable (Anderson & May 1991; Ma & Xia 2009). In such a setting, computer simulation offers an alternative approach to model the course of the epidemic through time.

ABS is a relatively new simulation approach to modeling complex systems. In ABS a system is composed of interacting, autonomous entities called agents whose behaviors, as well as interactions with one another and the environment, are programmed as a set of behavioral rules. While discrete-event simulation (DES) is still more common in operations research, ABS introduces a new way to model complex systems. Such systems are characterized by the fact that their aggregate properties cannot be deduced by simply looking at how each component behaves, since the interaction structure itself is playing a crucial role. In comparison with the top-down modeling approach of DES (where a system is broken into its components governed by the global system’s logic), ABS follows a
bottom-up approach that allows for patterns, structures, and behaviors to emerge that were not explicitly programmed into the models, but arise through the agent interactions. In the case of epidemics, ABS models provide a flexible and powerful platform to depict emergent epidemic growth patterns over time and space, and provide credible bases for policy making when calibrated to actual epidemic data (Burke et al. 2006).

2.1 RA Model Formulation

Available resources are to be allocated to control the spread of the epidemic through healthcare interventions that may include vaccination, prevention, and treatment programs. Such programs can be classified into behavioral (e.g., counseling, partner notification), and non-behavioral interventions (e.g., immigration restrictions, screening) (Rauner & Brandeau 2001). A healthcare intervention can target epidemic parameters in a specific compartment or in an entire population. Associated with each intervention is a production function describing the relationship between the amounts invested in the intervention and the future outcomes of the epidemic model (Brandeau 2004). In micro-economics, a production function specifies the output of a firm for all combinations of inputs. For epidemic-control programs, outcomes may include the number of averted infections (INFs), health-adjusted life years (HALYs) gained, or any other aggregate measure of public health.

The effectiveness of an intervention is usually measured by changes in an epidemic parameter’s values rather than the outcome. For example, influenza antiviral effectiveness can be measured in terms of future changes in viral load and infectiousness of the infected individuals (Longini et al. 2004). In this sense, an intervention targets some parameters of the epidemic model and consequently affects the final epidemic outcomes. However, the exact relationship between the amount invested in a program and the future value of these outputs is often unknown. For this reason, some researchers modify the definition of the
production function to capture the relationship between the amount invested in an intervention and the level of some intermediate variables associated with the intervention that will eventually affect epidemic outcomes (Friedrich & Brandeau 1998; Richter & Brandeau 1999; Brandeau et al. 2003). We refer to such functions as efficacy functions.

In this section, we formulate an epidemic RA problem based on (Zaric & Brandeau 2001) for interacting populations and interventions. All notations are summarized in Table 1. For convenience, we use capital letters for all global parameters that form the RA problem structure, and use their corresponding small letters for the running indices. We also use capital letters for the remaining quantities that are random variables (see Table 1). We consider a compartmental model of an epidemic existing among \( L \) populations where each population is divided into \( M \) health states. In order to model the heterogeneous structure of populations with regard to additional variables (e.g., age structure, risky behavior, socioeconomic status), each compartment is further divided into \( K \) sub-groups.
Global Parameters (Running Indices)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>L (l)</td>
<td>number of populations</td>
</tr>
<tr>
<td>M (m)</td>
<td>number of health states</td>
</tr>
<tr>
<td>L (k)</td>
<td>number of sub-groups</td>
</tr>
<tr>
<td>T (t)</td>
<td>length of time horizon</td>
</tr>
<tr>
<td>Π (g)</td>
<td>number of epidemic parameters</td>
</tr>
<tr>
<td>N (n)</td>
<td>number of interventions</td>
</tr>
<tr>
<td>O (o)</td>
<td>number of objective functions</td>
</tr>
<tr>
<td>B</td>
<td>total available budget</td>
</tr>
</tbody>
</table>

RA Decision Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>vₙ</td>
<td>amount invested in intervention n</td>
</tr>
<tr>
<td>媒介</td>
<td>vector of investments values: (v₁, v₂, ..., vₙ)</td>
</tr>
<tr>
<td>ρₙ</td>
<td>investment ratio in intervention n</td>
</tr>
<tr>
<td>媒介</td>
<td>vector of investments ratios: (ρ₁, ρ₂, ..., ρₙ)</td>
</tr>
</tbody>
</table>

Other Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ωₔ</td>
<td>sample space for g_th epidemic parameter</td>
</tr>
<tr>
<td>Ω</td>
<td>global sample space for a set of epidemic parameters</td>
</tr>
<tr>
<td>Pₔₕ (t; Ωₔ)</td>
<td>epidemic parameter g of general form, at time t, over Ωₔ</td>
</tr>
<tr>
<td>P(t; Ω)</td>
<td>set of epidemic parameters at time t, over Ω</td>
</tr>
<tr>
<td>Wₔₕ (t;媒介; Ωₔ)</td>
<td>g_th efficacy function: value of parameter g over sample space Ω, at time t, given investment媒介</td>
</tr>
<tr>
<td>媒介 (t;媒介;Ω)</td>
<td>vector of efficacy functions over sample space Ω</td>
</tr>
<tr>
<td>vₙMin</td>
<td>minimum limit for investment in intervention n: 0 ≤ vₙMin ≤ vₙ</td>
</tr>
<tr>
<td>vₙMax</td>
<td>maximum limit for investment in intervention n: vₙ ≤ vₙMax ≤ B</td>
</tr>
<tr>
<td>dqₘ</td>
<td>disability-adjustment coefficient for individuals in health state m</td>
</tr>
</tbody>
</table>

Calculated Quantities

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xₙₖₙₗₚₙ (t;媒介(t;媒介;Ω))</td>
<td>size of sub-group k, in health state m, of population l, at time t, given the investment vector媒介</td>
</tr>
<tr>
<td>H₀媒介</td>
<td>objective function o of general form, given investment vector媒介</td>
</tr>
<tr>
<td>媒介</td>
<td>vector of RA objective functions</td>
</tr>
</tbody>
</table>

50
The RA problem is defined as finding the optimal investment of a limited budget $B$ among $N$ available interventions to optimize $O$ objective functions related to the spread of the epidemic over a time horizon of $t$ periods. Let $P_g(t;\Omega_g)$ (or $p_g(t)$) represent the $g$th stochastic (or deterministic) epidemic parameter at time $t$. The random variable $P_g(t;\Omega_g)$ is defined over the sample space $\Omega_g$. Sometimes parameters’ values in deterministic models are set to the mean of observed values, and the information on the variance is ignored. However, if the variance of the observed values of a parameter is high and the results are sensitive to that parameter, then the confidence in the results would be low. For example, in the context of an influenza pandemic (for more details see Section 4), let $P_g(t;\Omega_g)$ be the number of days a symptomatic patient spends in society before withdrawing to home, which could vary between, say, 0 to 4 days equiprobably. In a deterministic setting, the parameter might simply be assumed to have a fixed value of 2 days for everyone, so that $\Omega_g = \{2\}$ and $P_g(t;\Omega_g) = p_g(t)$ is deterministic. In the stochastic formulation, however, we let $\Omega_g = \{0, 1, 2, 3, 4\}$ and define $P_g(t;\Omega_g)$ as a stochastic parameter, such that $\Pr(P_g(t) = c) = 1$ for $c \in \Omega_g$. For convenience, we therefore choose the notation $P_g(t;\Omega_g)$ to represent a general form (of discrete/continuous stochastic or deterministic nature) for the $g$th epidemic parameter at time $t$, and let $P(t;\Omega) = \{P_1(t;\Omega_1), P_2(t;\Omega_2), \ldots, P_\Pi(t;\Omega_\Pi)\}$ be a set of $\Pi$ parameters describing the epidemic model at time $t \in \{1, 2, \ldots, t\}$ over sample space $\Omega$.

We define $v_n$ as the amount of investment in intervention $n = 1, \ldots, N$, and let $\vec{v} = (v_1, v_2, \ldots, v_N)$ be the vector of investments. Similarly, we define the investment ratio $\rho_n = \frac{v_n}{B}$, and let $\vec{\rho}$ be the vector of investment ratios. We assume that $v_n^{\text{Min}} \leq v_n \leq v_n^{\text{Max}}$, where $v_n^{\text{Min}}$ and $v_n^{\text{Max}}$ are the minimum and maximum allowable amount of investment in intervention $n$, so $0 \leq v_n^{\text{Min}}$ and $v_n^{\text{Max}} \leq B$. The investment amounts $v_n$ are the decision variables of the RA problem. Here, we consider the one-time RA problem, where the investments are
fully made at time zero and may have an immediate effect on the associated parameters. For more information about the dynamic RA problem, see (Zaric et al. 2002).

Associated with each intervention is an efficacy function that relates the amount invested in a program to the future values of parameters in \( P(t; \Omega) \). Let \( W_g(t; \bar{v}; \Omega_g) \), \( g = 1, \ldots, \Pi \), be the future value of the \( g \)th parameter with state space \( \Omega_g \) at time \( t \), after the investment of \( \bar{v} \) has been made. This function is used to model the effectiveness of available interventions on a parameter value. Based on the type of associated parameter, and the nature of the relationship between the interventions and the parameter, \( W_g(t; \bar{v}; \Omega_g) \) can be deterministic or stochastic. In practical cases, a closed form of such a function is seldom known and the relationship is estimated through clinical trials, e.g., an antiviral’s (intervention) effectiveness on reducing the infectiousness of a patient (parameter) can be measured and is usually stated as a confidence interval. While most prior models used a point estimator of such parameters, inclusion of uncertainty can provide a more realistic framework that generalizes previous models, and conveys information on risk. We let

\[
\bar{W}(t; \bar{v}; \Omega) = (W_1(t; \bar{v}; \Omega_1), W_1(t; \bar{v}; \Omega_1), \ldots, W_{\Pi}(t; \bar{v}; \Omega_{\Pi}))
\]

be the vector of efficacy functions mapping the investment amounts into the future epidemic parameter values, \( R^N \rightarrow R^\Pi \), in the sample space \( \Omega \).

Let \( X_{\text{inf}}(t, \bar{W}(t; \bar{v}; \Omega)) \) be the size of sub-group \( k \) in health state \( m \) of population \( l \) at time \( t \), given the efficacy function vector \( \bar{W} \) and investment vector \( \bar{v} \). This is denoted with a capital letter, as it is a function of (possibly) stochastic epidemic parameters through time. Initial group sizes \( X_{\text{inf}}(0, \bar{W}(0; \bar{v}; \Omega)) \) are known (or estimated) for all sub-populations at time 0. Let \( H_o \) be an objective function of general form in the RA problem. The \( H_o \) value usually depends on the sub-groups’ sizes \( (X_{\text{inf}}) \) through time (e.g., INFs is a function of the
infected sub-groups’ sizes), which in turn are influenced by the investment strategy; and so
the notation $H_0(\bar{\nu})$ is used. We let $\overline{H(\bar{\nu})} = (H_1(\bar{\nu}), H_2(\bar{\nu}),\ldots, H_0(\bar{\nu}))$ be the vector of such
objective functions, and define the RA problem as:

$$\text{RA: Optimize } \overline{H(\bar{\nu})}$$

s.t. $\sum_{n=1}^{N} v_n \leq B,$

$$v_n^{\text{Min}} \leq v_n \leq v_n^{\text{Max}}$$

Assumptions regarding the type and definition of the objective functions differ among
studies based on the scope and goals of policymaking. Longini et al. (2004) study the
effectiveness of antiviral agents in the control of influenza and consider two measures of
intervention effectiveness, the average overall effectiveness and the epidemic prevention
potential (Longini et al. 2004). The average overall effectiveness is 1 minus the ratio of the
average attack rate (AR) in the intervention populations to the average attack rate in the
nonintervention populations. The illness attack rate is the cumulative incidence of infection
in a group of people observed over a period of time during an epidemic. A more general
type of objective function, widely used in medical decision making, is the population’s
HALYs measured through disability vs. quality adjusted life years (DALYs vs. QALYs).
Assuming the disability-adjustment coefficient for life years lived by individuals in different
health states, such that the healthier individuals have lower disability ($d_{\text{susceptible}} \leq d_{\text{infected}}$),
DALYs is computed as the sum-product of the total number of people in each health state
and the associated DALYs coefficients, over the time period $t$: 

53
The goal of RA is to minimize the total DALYs experienced after investing in available interventions. While DALYs is a measure of disease burden, the QALYs is a measure of health gain, and is computed in an inverse fashion.

### 2.2 Analysis of the RA Problem

The RA problem for epidemic control is complex, and differs in a number of significant ways from traditional RA problems. Examples of such complexities include:

- nonlinear and dynamic epidemic growth rate over time,
- the relationship of investments and the epidemic parameters modeled by linear or nonlinear, deterministic or stochastic, independent or interacting efficacy functions,
- the study time horizon affecting the optimal solution in non-stationary processes,
- additional assumptions regarding disease pathology (e.g., natural history, means of transmission), social structure (e.g., nonhomogeneous mixing groups, dynamic population size), and other economical, political, and social objectives of the study (e.g., achieving equity among groups or programs, focusing on under-served populations, restricted access to certain programs).

Researchers have addressed the epidemic RA problem using a variety of approaches including linear and integer programming, optimal control, equilibrium analysis, simulation, and numerical analysis. Brandeau (2004) reviews the literature and discusses the shortcomings of the proposed models in addressing different aspects of epidemic RA.
complexities. For example, optimal-control analysis is often restricted to problems with a single intervention, the equilibrium approach is limited by the assumption of a sufficiently long time horizon for equilibrium to be reached, and simulation applications have usually limited their analysis to consideration of a small set of given and fixed alternatives.

Recent studies propose a variety of optimization approaches to overcome these limitations, usually employing either a very simple epidemic model or an approximation of a more complex model. A recent analysis considers allocation of epidemic-control resources to multiple non-interacting populations (Brandeau et al. 2003). The epidemic in each population is described by a simple epidemic model (susceptible and infective), and interventions target the sufficient contact rate of the disease. Two objective functions, QALYs and INFs, are considered and analytical results are developed for a sufficiently long time horizon. In another work, a more comprehensive optimization framework is developed that allows for interacting populations and interacting prevention programs. The problem is analytically intractable even for simple epidemic models, and several heuristics are presented to approximate the optimal solutions (Zaric & Brandeau 2001). Kasaie et al. (2010) propose a simulation-based technique, and show the validity of their approach using a simple RA problem with a closed-form mathematical solution’s providing a benchmark.

Here we address a rather general class of one-time epidemic RA problems, and propose a simulation-optimization approach to find the best investment strategy. We relax the prior restrictive assumptions regarding the number of interventions (let \( N \geq 1 \)), form and independency of efficacy functions (\( \overline{W}(t; \tilde{v}; \Omega) \) is not restricted in form), length of the time horizon (let \( t \geq 0 \)), and deterministic epidemic parameters (\( P(t; \Omega) \) can be stochastic or deterministic). The freeform RA problem, as we refer to it for the rest of this paper, has
minimal restrictions regarding the underlying epidemic model and associated parameters (e.g. a non-Markovian model of several compartments associated with stochastic or deterministic parameters is allowed), as well as the definition of the RA problem and form of efficacy functions (e.g., a problem concerning several interacting interventions, with nonlinear efficacy functions addressing multiple parameters). While the suggested approach is a combination of existing simulation and optimization techniques, we propose a novel application of these tools to address the freeform epidemic RA problem. In the following section, we present the suggested simulation-optimization methodology, and illustrate the implementation steps with a focus on the analysis aspect and optimization.

3 Methodology

The epidemic RA problem involves special complexities regarding epidemic modeling and RA structure that are highly resistant to standard analytical techniques, and make the models intractable (Section 2.2). The modeling complexities, however, can be easily addressed through a simulation model of an epidemic. Simulation models provide a flexible and powerful platform to model epidemics over time, and are credible bases for policy analysis when calibrated to actual epidemic data (Burke et al. 2006). The literature contains several instances of simulation models applied to the study of different infectious-disease epidemics, and models are still growing in popularity (Longini et al. 2004; Longini et al. 2007; Epstein et al. 2007; Grune-Yanoff, 2010; Chao et al. 2010).

While a simulation model provides a realistic representation of a system’s current behavior, it can be also used to study the system’s future behavior under potential changes of model parameters. In this sense, an epidemic simulation model can be seen as a black box that models the effect of healthcare interventions in the RA problem, and represents the system’s future outcomes under each investment strategy (Figure 1). Coupling a simulation
model with statistical optimum-seeking techniques enables us to study the simulation outputs behavior in an effort to find the best setting of the model’s inputs. In the epidemic RA problem, we can modify the simulation model to incorporate the RA structure, and use simulation-optimization techniques to seek the best investment strategy. The methodology, therefore, consists of three main steps:

3.1 Step 1: Developing a Simulation Model of the Epidemic.

Different simulation techniques are applied to model epidemics and study the disease prevalence over time (Mellor et al. 2011; Atun et al. 2005; Brouwers 2005). Available techniques differ significantly in their modeling paradigm, and level of system representation (Rahmandad & Sterman 2008). For example, system dynamics follows a top-down approach, using feedback loops and stock and flows, and presents a global view of epidemic pattern and stock sizes over continuous time; while ABS uses a bottom-up
approach, defining the individuals’ (agents) attributes and behaviors and following their interactions, and presents the disease prevalence across time and space. In this sense,

ABS provides a more flexible framework to incorporate population heterogeneity, and enables the study of an epidemic’s outcomes at multiple levels, e.g., micro-level outcomes concerning each individuals’ health state such as DALYs, or macro-level outcomes concerning the population’s health such as INFs. With regard to such advantages, we adopt an ABS modeling paradigm for our study. Since our focus is not on the simulation modeling of an epidemic, but rather the analysis of such models and addressing the epidemic RA problem, we refer the reader to (Macal & North 2010; Bonabeau 2002) for more information on ABS characteristics and applications, and to (Sargent 2009; Troitzsch 2004; Klugl 2008) for additional discussion on verification and validation of simulation models.

In a typical epidemic ABS model, agents are humans whose characteristics and behaviors are modeled through attributes and behavioral rules. The natural population dynamics and disease transmission are modeled through agents’ interaction and the contact-network’s topology. The simulation model inputs include a set of parameters describing the population structure (e.g., population age structure, rates of birth/death in each compartment), as well as the epidemic system (e.g., viral load of infected people, length of infection/incubation period). We let \( \theta \) be a set of such input parameters to the simulation model, with no restriction regarding their form and values (Figure 1). The simulation outputs, denoted by \( \overrightarrow{f}(\theta) = (f_1(\theta), f_2(\theta),..., f_O(\theta)) \), are stochastic functions of the simulation input set \( \theta \). Each simulation output, \( f_o(\theta), o=1, .., O \), is defined with regard to the epidemic RA problem’s objective function \( H_o(\overrightarrow{v}) \), that is often described as a measure of population health, such as, DALYs, INFs, etc.
3.2 Step 2: Incorporate the RA Problem into the ABS Model.

The simulation inputs $\theta$ correspond to the epidemic parameters in $P(t; \Omega)$, whose values are in turn under the influence of healthcare interventions. This relationship is modeled through the vector $\overrightarrow{W}(t;\bar{v};\Omega)$ of efficacy functions that computes the future value of parameters after an investment Strategy $\bar{v}$ is selected. In order to incorporate the RA problem in the epidemic ABS model, we design an initial computational procedure (shown as Proc1: Parameter Translation in Figure 1) called at the beginning of our simulation code, and use the vector $\overrightarrow{W}(t;\bar{v};\Omega)$ to project the investment vector into the epidemic parameter-space, $R^N \rightarrow R^H$. The set of updated parameters ($\theta$) then inputs to the ABS model, which simulates the behavior of the epidemic system under the new setting, estimates the compartment sizes $X_{\text{inh}}(t,W(t;\bar{v};\Omega))$, and measures the influence of the investment strategy through simulation outputs in $f(\theta)$.

Moreover, we define the simulation outputs in correspondence to the RA problem objective functions, such that $f^o_v(\theta)$ computes the value of $H_v(\bar{v})$ at the $r^{th}$ simulation replication. Due to the stochastic nature of each simulation run, the objective function $H_v(\bar{v})$ is usually defined as the expected value (long-term average) of the corresponding output $H_v(\bar{v}) = E(f_v(\theta))$, which is estimated using $R$ replications of the simulation model under the same input setting (same investment strategy), $\hat{H}_v(\bar{v}) = \frac{\sum_{r=1}^{R} f^o_v(\theta)}{R}$. This computational procedure (shown as Proc2: Objective Function Estimation in Figure 1) is called at the end of each simulation run (covering $R$ replications).
3.3 Step 3: Apply Simulation Optimization to the Freeform RA Problem.

Iterative optimization methods are often used in conjunction with simulation models to search for designs with desired properties. Barton and Mechesheimer (2006) define the general simulation optimization problem as: “repeated analysis of the simulation model with different values of design parameters, in an attempt to identify best simulated system performance” (Barton & Mechesheimer 2006). In our problem, the simulation inputs can be traced back to the initial investment values in \( \bar{v} \), and the simulation outputs estimate the RA’s objective functions \( \bar{H}(\bar{v}) \), as shown in Figure 1. We therefore consider the simulation model as a black box, and define the optimization problem as repeated analysis of the epidemic ABS model with different investment values in \( \bar{v} \), subject to the budget constraint
\[
\sum_{n=1}^{N} v_n \leq B,
\]
in an attempt to identify the best values of estimated RA objective functions \( \bar{H}(\bar{v}) \).

Simulation-optimization strategies depend on the nature of the decision variables and objective function (model response). Based on our prior definition of the freeform RA problem, the decision variables \( v_n \) are deterministic, and the objective functions \( \bar{H}(\bar{v}) \) have general form. Barton and Mechesheimer (2006) review different strategies and divide them into two classes of methods involving discrete and continuous variables (Barton & Mechesheimer 2006). In the continuous case, further division leads to direct gradient methods when the objective function is differentiable, and metamodel methods for other cases. Stochastic gradient-based optimization can use efficient methods to estimate the gradient of the model response, without attempting to provide a global approximation to it. The stochastic optimization codes, however, are external to simulation, are often complex to implement, and can fail on stochastic responses with large variation. Metamodel
optimization methods use a fitted meta-model rather than the simulation response. A metamodel is a deterministic approximation function of the simulation response. Once a metamodel is in hand, optimization can be carried out using a sequential optimization procedure, known as response surface methodology (RSM) (Kleijnen 1998).

RSM is a metamodel-based optimization method that builds local approximations of a response, and uses a deterministic gradient-based optimization strategy to find a direction of improvement and follow it. While the simulation-optimization literature includes a wide selection of different optimization methods (Fu 2005), we choose RSM as our optimization technique due to the following advantages: 1. RSM is model-free, 2. it has a flexible framework with minimal restrictive assumptions regarding the variables’ and responses’ definition that allow us to address a general class of freeform RA problems, 3. RSM is well-established (using statistical experimental design) (Montgomery 2001), and 4. the optimization procedure involved is fairly straightforward to apply. One drawback however, is that RSM is not implemented in the simulation commercial packages, which requires us to code and implement it external to the simulation model. Another suggested drawback is the excessive use of simulation points in one area before exploring other parts of the search space, which is exacerbated for a large number of input variables. Van Beers and Kleijnen (2003) propose kriging techniques as a possibly more efficient way of carrying out this step, which is outside our scope (Van Beers & Kleijnen 2003).

**RSM Procedure for the Epidemic RA Problem:** The stepwise RSM procedure begins with specifying an initial vector of investments and running experimental designs to estimate the local response behavior at this point. First-order polynomials are often used through the initial RSM loops to identify the most promising direction of improvement:
where the $\beta_n$ s are the estimated regression coefficients, and $Y_o(\bar{v})$ is the polynomial fitted to $\hat{H}_o(\bar{v})$. The procedure continues by following the direction of improvement (steepest ascent in a maximization problem or steepest descent (SD) in a minimization problem) until no further improvement is observed, and then a new experimental design is implemented. The limited overall budget $B$ constrains the RA investments and may require alteration of step sizes along the improvement direction. Once the region of the optimum has been found, a more elaborate model, such as a second-order polynomial, may be used to provide a more realistic representation of the response behavior and locate the optimum:

$$Y_o(\bar{v}) = \beta_0 + \sum_{n=1}^{N} \beta_n v_n + \varepsilon, \quad \varepsilon \sim i.i.d. N(0, \sigma^2), \quad o = 1, 2, ..., O,$$

The optimization procedure is model-free, and treats the simulation model as a black box (Figure 1). This allows us to avoid the unnecessary complexities for integration of RSM into the ABS model, and to execute the optimization procedure external to the ABS model using available statistical packages, such as Design Expert (Design-Expert 2010), to design the experiments and analyze the metamodels, and general-purpose packages, such as Microsoft Excel VBA (Excel 2010) for interoperability between the simulation and optimization models.
4 Illustrative Case

In this section, we provide an illustrative case of an RA problem example in the control of an influenza pandemic, a fairly complicated problem for which no analytical solution is available. We aim to design a general case of the RA problem that has the complexities regarding the population structure and the epidemic system, as discussed in Section 2.2. We adopt a respected simulation model of influenza from the literature, which has been previously verified and validated, and define the RA problem in the context of an influenza pandemic. We then apply the suggested simulation-optimization approach to a specific situation, and illustrate the implementation steps.

4.1 Simulation Model of Influenza

We adopt an ABS model of influenza (FluTE) and modify it to incorporate the associated RA problem (Chao et al. 2010). FluTE is written in C/C++ and the source code is available at http://www.csquid.org/software. The model simulates stochastic spread of influenza across an age-structured population of individuals interacting in known contact groups. Incorporating a sophisticated natural history of influenza and more realistic intervention strategies, FluTE has been used in a number of policy-making studies to compare the effectiveness of intervention strategies (Longini et al. 2004; Longini et al. 2005; Longini et al. 2007).

We adopt an instance of FluTE for a population of 2,000 people based on (Longini et al. 2004). The population is randomly generated using the U.S.-wide family-size distribution from the 2000 Census. The model has been calibrated so that outcomes are consistent with the 1957/1958 Asian A(H2N2) and the 2009 pandemic A(H1N1) influenza viruses. We adopt the default values of all the model’s parameters and probabilities and modify them for
the RA problem (Kasaie & Kelton 2013a). We now briefly discuss the model’s structure and refer the reader to (Chao 2013), and (Longini et al. 2004) for more information and details.

**Population Structure and Social Contacts:** Each person (agent) in the model is characterized by a set of attributes regarding the age group (five age groups), memberships to various mixing groups, disease status, etc. The population is organized as a hierarchy of increasingly large but less-intimate mixing groups, from households, household clusters, neighborhoods, and the community. Including additional mixing groups, such as daycare centers, schools, and work places, creates a realistic contact network for disease transmission. The disease transmits itself through direct contact of infectious individuals with susceptible ones. The contact probability of two individuals in the same mixing group is the probability that they will have sufficient contact for transmission during a time step. The probabilities vary among different mixing groups and age levels. The simulation runs in discrete time for 180 days, with two time steps per simulated day to represent daytime and nighttime social interactions.

**Natural History of the Disease:** Influenza is introduced by randomly assigning 12 initial infectious persons. The compartmental model of influenza consists of six disease stages. Infected individuals first go through a latent period when they are not infectious and do not have influenza symptoms. This is followed by the infectious period during which people may develop influenza symptoms with certain probabilities. Symptomatic individuals are twice as infectious as asymptomatic people and may withdraw to home after 0 to 2 days equiprobably (where they interact only with their households). Six days after infection, an individual recovers and is no longer susceptible.

**Simulated Interventions:** Vaccination and antiviral treatments are the two primary pharmaceutical interventions. Interventions target three parameters of the epidemic model for each person: the probability of becoming infected, the conditional probability of
becoming ill given infection, and the probability of transmitting the infection. Vaccines do not reach full efficacy immediately and their protective effects may gradually increase over several weeks. In our example, vaccines take 30 days to reach maximum efficacy, with the efficacy increasing exponentially starting the day after the vaccination, and require a boost on day 21. Antiviral agents are used for treatment of existing cases, and for prophylaxis of susceptibles. A single course of antiviral agents is enough for ten days of prophylaxis or five days of treatment. Unlike vaccines, the protective effects of the antiviral agents last only as long as they are being taken. Interventions can be administered under different strategies. We consider a reactive strategy of mass vaccination to vaccinate all people once the epidemic is detected. Moreover, once a case is ascertained (with a 70% ascertainment chance and a one-day initial delay), the individual is treated with antiviral agents, and the individual’s household members are also given a course of antiviral prophylaxis (the HHTAP strategy).

4.2 The RA Problem in Control of an Influenza pandemic

We consider the described model structure with \( t = 180, L = 1, M = 6 \), and arbitrarily choice of \( K \) (Section 2.1) and assume two objective functions for this RA problem: the total influenza AR, and the total value of DALYs. The AR value is determined at the end of each model replication by the cumulative incidence of influenza infection throughout the population (over time \( t \)). The DALYs value is computed through (1), where \( dq_1 \) and \( dq_2 \) are assumed to be the disability coefficients associated with the infected and symptomatic health states with nonzero values, such that \( dq_2 \geq dq_1 > 0 \), and \( dq_m = 0, m = 3, 4, 5, 6 \).

4.2.1 Calibration, Sensitivity Analysis, and Experiments

We consider a simple case with two available interventions targeting two model parameters: the total order of vaccine units \( P_1(t; \Omega_1) \), and available antiviral doses \( P_2(t; \Omega_2) \), where \( \Omega_1 \),
\( \Omega_2 \in \mathbb{N} \), and have no upper constraint for the possible orders’ amount. This simplified case is used to calibrate the RA objective functions, study the interventions’ efficacy, and analyze the response behavior in the vicinity of the optimum. We then expand this problem in Section 4.3 to a more general case with six interventions.

**DALYs Function Calibration:** We assume a total budget of \( \$40,000 \) (\( B \)) to be invested at the beginning of the study. Let \( W_1(t; \bar{v}; \Omega_1) = \frac{v_1}{10} - 500 \) and \( W_2(t; \bar{v}; \Omega_2) = \frac{v_2}{40} \) be the efficacy functions associated with each parameter, \( P_1(t; \Omega_1) \) and \( P_2(t; \Omega_2) \), and let \( v_n \) be the investment amount in intervention \( n \), so that \( v_1 + v_2 \leq \$40,000 \). While the exact numerical values of parameters and function forms are chosen rather arbitrarily, we were guided by relevant studies on vaccine supply chains (Chick et al. 2008) and influenza antiviral production (http://www.flu.gov/vaccine/antiviral_use.html, accessed on October 12, 2011).

In order to set the numerical values of \( dq_1 \) and \( dq_2 \) in the DALYs function, we perform a sensitivity analysis of our models’ outputs for different combinations of these coefficients. Figure 2 shows a one-way sensitivity analysis of model outputs to different values of \( dq_2 \) with a fixed value of \( dq_1 \). Assuming a binding budget constraint \( v_1 + v_2 = \$40,000 \) (so that the whole budget is spent), the output values are plotted as functions of \( v_2 \).

![Figure 2: Sensitivity Analysis of AR (right) and DALYs (left) Values to dq2 With dq1 = 1.](image)
As expected, AR is insensitive to $dq_2$, with minor fluctuations due to randomness. The AR values in this case are very close and the plots (for $dq_1$, $dq_3$, and $dq_4$) are essentially on top of each other. On the other hand, $dq_2$ affects the shape and value of DALYs; with higher $dq_2$ the function is more convex and is higher. A similar effect is observed with $dq_1$ (Kasaie & Kelton 2013a). Since the horizontal location of the DALYs function is fixed, changes in the function’s convexity do not affect the location of the minimization valley. Thus, we don’t expect the $dq_i$ coefficients to affect the optimal solution. Conventionally, we choose $dq_1 = 1$ and $dq_2 = 2$, which agrees with our prior influenza model’s assumption in Section 4.1 (symptomatic individuals are twice as infectious as asymptomatic people).

**Two-Factor Experiments:** To gain initial insight into optimal allocation of resources and behavior of objective functions, we study the interventions’ efficacy and analyze the two-factor RA problem under the current settings. Consider the following three scenarios: no investment ($v_1 = v_2 = 0$), complete investment in vaccination ($v_1 = 40000$, $v_2 = 0$), and complete investment in antivirals ($v_1 = 0$, $v_2 = 40000$). Figure 3 compares simulation results for the daily number of symptomatic cases under each strategy for 700 replications of the model (providing a relative precision of at worst 0.01 for all outputs).
The simulation output summary values are compared in Table 2. While the all-vaccination strategy slightly lowers the maximum number of symptomatic cases and diminishes the number of symptomatics late in the epidemic, the all-antiviral policy has a bigger effect on the peak number of symptomatics as well as a slight increasing effect late in the epidemic. Vaccine efficacy in this case is influenced by the model’s assumptions of daily production capacity (default of 50 doses a day) and initial delay (default of 10 days), which postpone the vaccine’s administration and consequently lower the overall efficacy.

Table 2: Objective Functions’ Values and 95% Confidence Intervals Under the Three Scenarios.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>AR</th>
<th>DALYs</th>
</tr>
</thead>
<tbody>
<tr>
<td>No investment</td>
<td>0.38 ± 0.001</td>
<td>9213 ± 44</td>
</tr>
<tr>
<td>Complete investment in vaccination</td>
<td>0.33 ± 0.002</td>
<td>5086 ± 64</td>
</tr>
<tr>
<td>Complete investment in antivirals</td>
<td>0.27 ± 0.001</td>
<td>4440 ± 34</td>
</tr>
</tbody>
</table>

On the other hand, the objective functions’ values in Table 2 show a dramatic difference in scale. Conventionally, we re-scale these values to a similar scale of 0 to 1; we consider the
values of objective functions under the no-investment strategy (the worst case) as the upper-bound values for these functions \( \text{AR}_{\text{max}} \approx 0.4, \text{DALYs}_{\text{max}} \approx 9300 \) and standardize \( \text{AR} \) and \( \text{DALYs} \) with regard to these values. We refer to the standardized functions as \( R_1s \) and \( R_2s \), i.e., \( R_1s = \text{AR}/0.4 \), and \( R_2s = \text{DALYs}/9300 \).

We finally consider the binding budget constraint \( v_1 + v_2 = $40,000 \), and address the RA problem in one dimension. Figure 4 shows the standardized output values for different values of \( v_1 \) with a step size of $5,000. Both functions show similar behavior with regard to investment strategies, and suggest an optimal valley around $15,000 to $20,000 investment in vaccination. Figure 5 (the main larger graphs for both right and left) show the output values for a step size of $1,000 in this interval. By decreasing the step, the functions’ behaviors start to disagree with each other, such that for investments of greater than $16,000, \( R_1s \) increases and \( R_2s \) continues to decrease. Assuming equal importance for both objective functions, we perform our final search in the vicinity of this point with a step size of $100 as shown in the smaller plots in Figure 5. The functions disagree until the sixth step, followed by a global minimum at \( v_1 = $16,100 \). We therefore choose the investment vector of \( \hat{\rho} = (0.4, 0.6) \), as the near-optimal investment ratio that provides an average value of \( \text{AR} = 0.25 \) and \( \text{DALYs} = 3206 \) in our problem.
We expand the two-factor problem to a more general and realistic case with six interventions in the context of an influenza pandemic, as shown in Table 3. Vaccine production involves several operational challenges (Chick et al. 2008), like virus antigenic drifts that require annual reformulation of seasonal vaccines. Moreover, the production capacity for pandemic vaccine depends on available capacity for seasonal vaccine (Kieny & Fukuda 2008), and involves further challenges regarding stockpiling, administration, and distribution of vaccines. We model such constraints through the first three interventions’
targeting the vaccine’s total order, initial order delay, and daily administration capacity ($P_g(t, v; \Omega_g), g = 1, 2, 3$). The first intervention models the investment in vaccine order (possibly from another country) for the whole time horizon; with a rather large (relative to the total budget) initial cost of $5,000 for placing and processing the order. We assume no shortage in the amount of available vaccines, so that any order is met. Once the order is made, vaccines become available with a rather low unit price of $10. The second intervention models the initial delay in acquiring the ordered vaccines (possibly due to order processing, and delivery delay), and is defined as a step function. The default setting takes 10 days for the ordered vaccines to become available, which can be reduced to a minimum of 5 days (with a $5,000 cost for reducing each day). The third intervention models daily administration and distribution capacity with a default value of 50 units per day. The daily capacity can be increased by 50 units for every $2,000 and the capacity is assumed to be 300 units per day.
Table 3: RA Interventions and Associated Parameters.

<table>
<thead>
<tr>
<th>Intervention</th>
<th>Associated Parameter</th>
<th>Efficacy Function</th>
<th>$v_{Min}$</th>
<th>$v_{Max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>$P_{g}(t, \tilde{v}; \Omega_n)$</td>
<td>$W_n(t, \tilde{v}; \Omega_n)$</td>
<td>$v_1/10$</td>
<td>4000</td>
</tr>
<tr>
<td>1</td>
<td>Total available vaccine (units)</td>
<td></td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Vaccine's order initial delay (days)</td>
<td>$10 - \frac{v_2}{5000}$</td>
<td>0</td>
<td>2500</td>
</tr>
<tr>
<td>3</td>
<td>Vaccine's daily administration capacity (units)</td>
<td>$50 + \frac{v_3}{5000}$</td>
<td>0</td>
<td>1000</td>
</tr>
<tr>
<td>4</td>
<td>Total available antiviral (doses)</td>
<td>$\frac{v_4}{40}$</td>
<td>0</td>
<td>4000</td>
</tr>
<tr>
<td>5</td>
<td>Antiviral daily administration capacity (doses)</td>
<td>$100 + \frac{v_5}{2000}$</td>
<td>0</td>
<td>1600</td>
</tr>
<tr>
<td>6</td>
<td>New cases ascertainment delay (days)</td>
<td>$7 - \frac{v_6}{1500}$</td>
<td>0</td>
<td>9000</td>
</tr>
</tbody>
</table>

In contrast to influenza vaccine where availability and supply at the onset of a pandemic cannot be predicted, antiviral medications can be stockpiled and availability is assured. So we assume no initial investment cost for antivirals, but a higher per-unit cost of $40 in the fourth intervention. We address the antiviral daily administration capacity $P_{5}(t, \tilde{v}; \Omega_5)$ in the fifth intervention. The function has a form similar to the second intervention but with a higher initial and maximum daily value due to treatment availability, and a similar step function of $2,000 for each 50 doses increased. The reactive vaccine strategy begins at the onset of the epidemic, and antiviral HHTAP strategy is triggered with the ascertainment of new cases. To investigate the effect of initial-response delay, we include the sixth intervention parameter $P_{6}(t, \tilde{v}; \Omega_6)$ as an initial delay of a week for reactive response; future investments can reduce the delay to a minimum of three days.

Our choice of linear and step-function forms for the efficacy functions is based on the literature and affects neither the generality of the proposed framework nor the analysis.
procedure; see Kasaie et al. (2010) for an epidemic RA problem with a fairly complex nonlinear efficacy function. While the efficacy functions in Table 3 do not incorporate explicit dependencies among interventions, the inherent nature of the associated parameters in the epidemic model poses several dependencies on the programs’ efficacy and eventually the investment amounts in each intervention. For example, choosing no investment in the first intervention results in no vaccine availability for the whole period, and consequently cancels out the effect of the second and third interventions, so that \( v_2 + v_3 \leq v_1, \ v_n \geq 0 \). A similar indirect dependency exists among the fourth and fifth interventions for the total order of antiviral doses, as well as the first, second, and sixth interventions regarding the response delay, so that \( v_5 \leq v_4 \).

### 4.4 Analysis of the Expanded RA Problem

The optimization problem is:

\[
\begin{align*}
\text{Obj:} & \quad \text{Min R1s}(\bar{v}) \ & \text{Min R2s}(\bar{v}) \\
\text{s.t.} & \quad \sum_{n=1}^{6} v_n \leq 40,000, \quad v_2 + v_3 \leq v_1, \quad v_5 \leq v_4, \\
& \quad v_{n \text{Min}} \leq v_n \leq v_{n \text{Max}}
\end{align*}
\]

Here the \( v_{n \text{Min}} \) and \( v_{n \text{Max}} \) are the lower and upper bound limits for investment values in Table 3. The investment values \( v_n \) are deterministic decision variables in the optimization model. The objective functions, the R1s and R2s, are unknown stochastic, continuous functions of decision variables, estimated as expected values of the associated ABS model outputs (AR and DALYs). The optimization problem in (2), however, is a multi-objective problem, and requires further assumptions regarding the preference of each response in the optimization logic. The literature proposes the use of graphical techniques (overlaid contour plots),
constraint optimization, or a desirability function for such problems (Montgomery 2001). We adopt the last approach and define a desirability function \( d_o \) associated with each objective function \( H_o(\bar{v}) \) such that \( d_o = 1 \) if \( H_o(\bar{v}) \) is at its goal, and \( d_o = 0 \) if \( H_o(\bar{v}) \) is outside of the accepted region. The optimization problem is now stated as maximizing the overall desirability, \( D = (d_1 \times \ldots \times d_O)^O \), where \( O \) is the number of responses.

With these assumptions, we now present our analysis of the expanded model 2. We determine the required number of simulation replications to provide a satisfactory estimate of the objective functions (discussed next), and begin our optimization analysis by performing a global experimentation in Section 4.4.1. This experiment provides insight on global behavior of responses, and proposes a candidate starting point for the RSM procedure. In the following sections, we follow the RSM procedure, and repeat the analysis for three different initial points that suggest a same investment strategy.

**Determining the Required Number of Simulation Replications:** We use an average value of the simulation outputs over \( R \) replications as the point estimator of objective functions, \( \hat{DALYs}(\bar{v}) \) and \( \hat{AR}(\bar{v}) \), where number of replications \( R \) is tuned to provide a sufficiently precise confidence interval half-width for these estimates. Since our primary interest is in differences between responses across different input-parameter combinations, we tried properly synchronized common- random numbers (CRN) as a variance-reduction technique to reduce the required number of runs. Following the guidelines in (Law & Kelton 2000), we identified 46 sources of random variates in the model and assigned a separate random-number stream to each of them. The synchronization logic and numerical results of our analysis are provided in Kasaie and Kelton (2013). Comparing output variability using CRN, and using independent sampling (IS) throughout, we observed a significant reduction in the variability of differences in outputs using CRN. This reduction is a result of induced
positive correlation among design points. The CRN approach, however, has no effect on the variances of the individual outputs, so that the required number of runs to maintain a specific relative precision for each design point is similar under both strategies. Moreover, the induced positive correlation violates the independence assumption typically made in the analysis of experimental designs and the RSM technique, and brings into question the statistical validity of fitted polynomials using ordinary least square methods. One remedy is the use of more elaborate methods of generalized least squares for dependent observations, but this in turn requires additional statistical-software resources and overcomplicates the analysis procedure. With regard to these issues and the satisfactory speed of the simulation model under completely independent sampling in order to achieve acceptable precision, we decided to adopt the simpler and statistically “cleaner” IS approach.

<table>
<thead>
<tr>
<th>Relative Precision</th>
<th>DALYs</th>
<th>Number of Replications</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.01</td>
<td>1236</td>
</tr>
<tr>
<td>0.01</td>
<td>0.02</td>
<td>926</td>
</tr>
<tr>
<td>0.02</td>
<td>0.01</td>
<td>1225</td>
</tr>
<tr>
<td>0.02</td>
<td>0.02</td>
<td>310</td>
</tr>
<tr>
<td>0.03</td>
<td>0.03</td>
<td>138</td>
</tr>
</tbody>
</table>

We next perform an initial experiment for five random points, and estimate the required number of runs using the maximum of two approximation techniques, i.e,

\[ R = \max \left[ z_{\alpha/2} \frac{h^2}{s_0}, R_0 \frac{h^2}{h_0} \right], \]

where \( R_0 \) is the initial number of replications, and \( s_0 \) and \( h_0 \) are the standard deviation and desired 95% confidence-interval half-width interval around the mean response value (Law & Kelton 2000). From these initial \( R_0 \) runs, we choose the
maximum required number to provide the desired relative precision, \( e = \frac{h}{V} \), for each output at each point, and selected the maximum value as in Table 4. We choose a minimum of 400 replications for the initial experiments to provide a relative precision of at worst 0.02 for both responses. In the later steps of analysis, we increase this number to 1000 replications to ensure a relative precision of at worst 0.01 for DALYs, as the leading response in the optimization procedure.

**Global Experiment: A Mixture Design**

Mixture experiments apply when the factors are components or ingredients of a mixture and consequently, their possible values are related to each other. In constrained mixture designs, the factors are subject to additional constraints on their upper/lower bounds and their feasible combinations. Our investment values \( v_1, \ldots, v_6 \) are subject to a global budget constraint and are individually bounded to certain ranges as well. We therefore use a mixture experiment and a D-optimal constrained mixture design with 600 replications to achieve at worst 0.02 relative precision. The experiment results are used to fit linear metamodels to each response that approximates the response-surface behavior over the whole lattice. We use the desirability approach with equal weights to estimate the optimal minimizing solution. Figure 6 shows the Design-Expert (Design-Expert 2010) triangular mixture graph for desirability values in terms of the first three investment ratios (shown on each lattice vertex), i.e., \( \bar{\rho} = (v_1, v_2, v_3, 0.33, 0.00, 0.00) \). The three factors are bounded in a triangular lattice formed by the binding budget constraint so that the sum of the investment ratios equals one. The shades represent the desirability values across the feasible area of the lattice, and the optimal investment is numerically estimated at \( \bar{\rho} = (0.49, 0.14, 0.04, 0.33, 0.00, 0.00) \) with a desirability of 0.95. To read off the
coordinates, one can draw perpendicular lines to triangular altitudes that show the proportion of investment associated with each vertex (Kasaie & Kelton 2013a).

![Figure 6: Desirability Values in Mixture Design.](image)

### 4.4.1 Experiment 1

We use the solution suggested by the global mixture design as the initial point for the RSM procedure and begin our search for the optimal solution. The factor levels around the center point are defined such that they at least cover one step value in the associated piece-wise efficacy function, e.g., the $v_3$ levels around the initial point ($1,680$) are set to $0$ and $20,000$ corresponding to the step value of the third efficacy function. Moreover, we use the binding budget constraint to reduce the dimensionality to five, and define the fourth
investment value as a function of other investments’ amounts, i.e., \( v_4 = 40,000 - (v_1 + v_2 + v_3 + v_5 + v_6) \).

The optimization procedure and numerical results are detailed in (Kasaie & Kelton 2013a). The initial experiments use simpler designs with a low number of replications to fit a first-order polynomial to each response. Using the fitted metamodels, significant factors are identified, and the search continues in the direction of SD. Our priority in selecting the improvement direction is minimizing the DALYs function, while controlling for the behavior of the AR function. Similar to the two-factor case in Section 4.2.1, the global objective functions’ behavior is often similar and behaviors disagree only at small scales.

The significant factors at the end of Experiment 1 are \( v_1 \) and \( v_3 \), while \( v_4 \) is a binding variable and the other three factors (\( v_2, v_5, v_6 \)) are insignificant. We now face various strategies to select the next steps along the SD direction. The first traditional strategy is to fix insignificant factors’ value at the design center point, and change the others with respect to the desired step size, i.e., fixing \( v_2 = 7500, v_5 = 1000, \) and \( v_6 = 750 \), and changing \( v_3 \) and \( v_1 \) along the SD direction. A problem with this approach, however, is the low number of feasible steps along the improvement path before reaching the global-budget and factor-limit constraints. We therefore suggest a second strategy in which we set the insignificant factors’ values to their lower design limit, and allocate the released budget to other significant factors such that \( v_2 = 5000, v_5 = 0, \) and \( v_6 = 0 \). Figure 7 compares the SD path for DALYs values under the two strategies. Releasing the extra resources not only increases the number of feasible steps, but also improves (lowers) the average value of the response. Similar behavior is observed with the AR function.
Using the second strategy, we continue our search along the SD direction for 20 steps, until the improvement stops (Figure 8, left graph). Both responses have a diminishing negative slope until the 15th step, and start increasing afterward. So we choose this point as the initial point of the second RSM loop and design a new experiment. Following a similar procedure for the second RSM iteration, we fit a linear metamodel to each response and follow the proposed improvement path. For this loop, the only significant factor is $v_1$ with a step size of $500$, and the other insignificant factors are fixed at their lower design limits, i.e., $v_2 = v_5 = v_6 = 0$, $v_3 = 8,000$, and $v_4 = 40,000 - v_1 - v_3$. Figure 8 (right graph) shows the output values along the SD path for the second loop in which the improvement ceases at the 8th step. In the third loop, we design a D-optimal experiment with three remaining factors such that the overall budget constraint holds at equality. We reduce each factor’s level to a $500$ interval around the center point, and increase the number of replications to 1000 to achieve a worst case 0.01 relative precision for the $R2s$ (standardized DALYs) values. The design suggests a significant quadratic model for $R1s$ and a linear model for $R2s$, as shown in Figure 9. Using the desirability function, the best suggested point is $\hat{\rho} = (0.614, 0.000, 0.209, 0.177, 0.000, 0.000)$ with a desirability of $d = 0.771$. 

![Graph showing R2s over steps for two strategies](image)
Notice, however, that the suggested solution can still be improved with respect to the efficacy functions’ structure. In this case, the investment value of $v_3 = $8,363 can be fixed at $8,000 to release more resources for the other two interventions. So we reduce the dimensionality by removing the insignificant factors $v_2$, $v_5$, and $v_6$, and use a mixture design to fit a local higher-order polynomial to the responses. Figure 10 shows the desirability function with respect to $v_1$ and $v_4$ jointly with a fixed value of $v_3 = $8,000. Using a cubic fit on $R1s$ and a quadratic fit on $R2s$, Figure 10: the final suggested optimal point to maximize
the desirability function (minimize both responses with equal weights) is 
\( \bar{\rho}_1 = (0.632, 0.000, 0.200, 0.168, 0.000, 0.000), \) estimating \( R_{1s} = 0.544 \) and \( R_{2s} = 0.079. \)

The suggested solution shows minor improvements from the last suggested point and corresponds to the values of \( AR = 0.218 \) and \( DALYs = 735 \) in the original RA problem.

![Design-Expert® Software Component Coding: Actual Desirability](image)

**Figure 10:** Desirability Function for R1s and R2s With Two Factors.

### 4.4.2 Further Experiments

An important concern in applications of RSM and other similar heuristics is lack of a statistical guarantee for finding the global optimum. To address this problem, the literature proposes a multi-start technique, or multiple searches from the same starting point, to verify the suggested solution. So we performed two additional experiments with different starting points. The first experiment used the best design point found in the global mixture, and the
other experiment chose a random initial point such that it covers at least one level of each intervention. The analysis specifications and numerical results for each experiment are provided in (Kasaie & Kelton 2013a). We followed an optimization procedure similar to Experiment 1 and used the desirability approach to estimate the optimum. The suggested points in both cases ($\hat{\rho}_2 = (0.625, 0.000, 0.208, 0.167, 0.000, 0.000$ and $\rho_3 = (0.600, 0.000, 0.200, 0.200, 0.000, 0.000)$) are located in the vicinity of the prior solution presented above, which verifies our prior analysis. In conclusion, the suggested optimal investment strategy results in a 0.43% and 0.92% reduction of corresponding AR and DALY values in the original model (with no investment), and allocates the majority of funds to the total vaccine’s order (about 60%) and its administration capacity (about 20%). The current suggested strategy, however, depends on a variety of factors, and requires further sensitivity analysis of epidemic parameters and efficacy functions, to provide public-health recommendations for combating the influenza pandemic which is outside our scope.

5 Discussion and Conclusion

We have considered a class of freeform epidemic RA problems where a fixed budget is to be allocated among competing healthcare interventions to achieve the best health benefits, and proposed a simulation-optimization framework to address the problem.

In order to arrive at a more realistic representation of an epidemic’s complexity, we used a simulation model of an epidemic that represents the course of disease prevalence over time and space. We adopted the ABS modeling technique for its advantages over other available simulation paradigms in the context of the social sciences and epidemics. ABS allows us to model a variety of interventions, affecting an individual’s or a population’s parameters, and to measure their consequent effects at both the micro and macro levels. While simulation
modeling is not the focus of this work, we emphasize the importance of this step to provide reliable models that are verified and calibrated to available public health data.

Incorporating the epidemic RA problem into the epidemic ABS model, we applied RSM as an optimum-seeking technique to find the best investment strategy. The optimization logic was external to simulation model and treated it as black box. In the illustrative case example, we used Design-Expert to design the RSM experiments and analyze the results in each loop. We modified the simulation code to include the additional computational procedures (Figure 1), and used Microsoft Excel VBA for interoperability, e.g. managing the experiments’ execution, feeding the simulation inputs to FLuTE, and the generated simulation outputs to Design Expert, at each step. We chose RSM for it’s advantages over other optimization techniques as discussed in Section 3.3. A drawback, however, is lack of a guarantee for global optimality of the suggested point, and low computational efficiency of the search method for problems with many variables. We addressed this by fortifying our procedure with an initial global mixture design that suggests the RSM initial point, and performing multiple searches using different starting points.

The stochastic nature of computer simulation introduces realistic uncertainty into the model’s predictions, which in turn requires additional computational effort to provide precise results. We studied alternative random-number sampling approaches including IS and CRN, and realized the tradeoff between the cost of simulation runs and the cost of more elaborate analysis techniques. In the illustrative case, the number of simulation replications (to provide a relative precision of 2% at worst for all responses) induced a low computational cost, and lead to our choice of the statistically “cleaner” IS technique. For more expensive (or slower) simulation models, however, the required number of runs may not be feasible, and the tradeoff might tilt in favor of CRN (or other variance-reduction techniques) to reduce the required number of runs.
We have proposed and demonstrated novel application of multiple tools and techniques to a class of freeform epidemic RA problems with minimal modeling restrictions or oversimplifying assumptions. The simulation-optimization approach is most suitable for realistic complex problems involving complicated structure or high levels of detail and uncertainty, that cannot be validly addressed through conventional analytical approaches. Once established, the epidemic simulation model can re-calibrated to new empirical data over time, and provide a sustainable platform for future system studies. Our focus is development of the suggested approach, and illustrating a typical application to a (self-defined) RA problem of complex structure, with an emphasis on optimization analysis. The suggested protocol, however, can be adapted to address a wide class of research questions in the context of different diseases epidemics, and provide public health advice and recommendations for policy makers.

Future work involves a realistic application to the co-epidemic of HIV-TB, involving a more comprehensive study of potential research questions and policy-making challenges. We will extend the work on simulation modeling, its calibration and validation, as well as the optimization analysis, enhanced with heuristics to increase efficiency.

6 Acknowledgment

We acknowledge and thank Dr. Margaret Brandeau for her interesting works on epidemic RA problem and valuable suggestions that impacted the research idea. We also thank the two anonymous referees for their feedbacks and helpful recommendations in improving the manuscript.
7 References


Essay 3: Resource Allocation for Controlling Epidemics:
Calibrating, Analyzing, and Optimizing an Agent-Based Simulation

Parastu Kasaie, W. David Kelton

Department of Operations, Business Analytics, and Information Systems, University of Cincinnati, PO Box 210130, Cincinnati, Ohio 45221-0130, USA.

Abstract

This is a companion paper to Kasaie and Kelton (2013), and provides an extended discussion on the calibration, analysis, and optimization of an agent-based simulation (ABS) model of an epidemic. The detailed information is presented for the illustrative case of a resource-allocation (RA) problem in the control of an influenza pandemic as described in Section 4 of Kasaie and Kelton (2013). The suggested protocol, however, can be adapted to address other instances of the RA problem in the context of other infectious diseases epidemics. Section 1 defines all the parameters and run conditions for the ABS model. In Section 2 we discuss calibration of the RA objective function (DALYs function) and related sensitivity analyses. An investigation of variance reduction for the ABS model is presented in Section 3. A short discussion on application of mixture design and the corresponding triangular output displays for this design is discussed in Section 4. Finally, Section 5 presents the numerical results of the RA problem using a response surface methodology optimization approach.

Key words: Epidemic Simulation modeling, Resource allocation, Simulation optimization, Response-surface methodology
1 FluTE Simulation Parameters

In this section, we present the parameters and run conditions for the ABS model of an influenza pandemic (FLuTE). The latest version of the FLuTE model, including the sample data files is accessible at http://www.cs.unm.edu/~dlchao/flute/. In our example, we use the sample data file One.txt including a population of 2000 people. We adopt the parameter values used in Chao et al. (2010), and specify the decision variables of the RA problem (Section 4.3 in Kasaie and Kelton (2013)). Table 1 shows a list of the parameters (parameter name, description, and assigned values) used in our configuration of the simulation model. The parameter’s descriptions are taken from the README text file accompanying the model; to emphasize that these descriptions are directly from that source, they are in italics. Once the settings are determined, the configuration file is generated in a *.txt format, and is read into the simulation model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Assigned Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>label: string</td>
<td>A name that is output in the summary file for the user's Run parameter</td>
<td></td>
</tr>
<tr>
<td></td>
<td>convenience (e.g., &quot;simulation1&quot;). With this value, the user can tag</td>
<td></td>
</tr>
<tr>
<td></td>
<td>output files from specific simulation runs.</td>
<td></td>
</tr>
<tr>
<td>Datafile: string</td>
<td>The prefix for the input data files names (e.g., one, Seattle, LA, One.txt USA).</td>
<td></td>
</tr>
<tr>
<td>logfile: integer</td>
<td>Number that indicates how often, in days, output is written to the Run parameter Log file. If 0, no Log file is generated. Default is 1 for daily output.</td>
<td></td>
</tr>
<tr>
<td>individualfile: binary</td>
<td>Specifies if an &quot;Individuals&quot; file should be generated. The file will Run parameter contain one row of simulation-specific data for each individual and can be very large. If 1, such a file is generated. The default is 0, no file generated.</td>
<td></td>
</tr>
<tr>
<td>beta: real</td>
<td>Transmission parameter (sometime known as Ptrans). This value 0.3 is multiplied by the inter-personal contact probability to determine the transmission probability between infected and susceptible people.</td>
<td></td>
</tr>
</tbody>
</table>
R0: real  Basic reproductive number, which is internally converted to beta.

runlength: integer  Number of days the simulation should run.  180

preexistingimmunitylevel: real  Those with pre-existing immunity have susceptibility reduced by preexistingimmunitylevel.

preexistingimmunitybyage: real[5]  Vector of five values representing the fraction of individuals in each age group with pre-existing immunity. Those with pre-existing immunity have susceptibility reduced by preexistingimmunitylevel.

defaultVESbyage: real[5]  Vector of five values representing the "VES" of individuals in each age group. A value of 0 indicates that the age group is fully susceptible and 1 indicates that it is completely immune.

prestrategy: string  Vaccination strategy before the epidemic occurs. Can take one of the following four values: none, prevaccinate, primeboostrandom or primeboostsame. The prevaccinate option vaccinates a fraction of the population (with 2 doses if required). The primeboostrandom option will give a percentage of the population one shot before the epidemic, then may or may not boost those same people after the epidemic has started. People vaccinated after the epidemic has started will be selected at random. In contrast, primeboostsame will boost the same people after the epidemic that were primed before the epidemic.

reactivestrategy: string  Vaccination strategy for when the epidemic is detected. Can take mass one of the following three values, none, tract, or mass. The tract option will vaccinate people in a specific tract once the threshold (i.e., responsethreshold variable) is met. The mass option will vaccinate all people in all tracts once the threshold is met.

vaccinationfraction: real, scalar  Fraction of people to vaccinate if a vaccination strategy is selected.

vaccinepriorities: integer[13]  A vector of 13 numbers, representing the vaccine priority for the 13 categories of individuals. A value of 1 indicates highest priority, 2 is the next-highest priority, etc. 0 indicates that this category is not prioritized to get vaccine.

Individuals of lower priority only get vaccine when those with higher do not need any. If a high-priority individual is vaccinated, lower-priority people can get vaccine until the high-priority person needs the boost. The categories, in order, are: essential workforce, pregnant women, family members of infants, high-risk preschoolers, high-risk school-age children, high-risk young adults, high-risk older adults, high-risk elderly, all preschoolers, all school-age children, all young adults, all older adults, all elderly. For example, to have individuals of all ages get
the same priority, we set all age-specific priorities to the same value: 0 0 0 0 0 0 0 1 1 1 1 1. To give adults lower priority we can do: 0 0 0 0 0 0 0 1 1 2 2 2.

antiviral doses: integer, scalar
Number of antiviral courses available at the beginning of the simulation. A single course is defined as 10 pills per person. Variable

vaccine doses: integer[2]
The vaccine ID followed by the number of vaccine doses available at the beginning of the simulation. A dose is defined as one shot Variable per person. If the vaccine is a split dose, (i.e., prime and boost), this variable must be twice the total number of doses. To specify 2,000,000 doses of vaccine 0, one would enter 0 2000000.

vaccine production: integer[runlength+1]
Vaccine ID followed by the number of vaccine doses that become available each day during the response. For example, a vector of 1 100 300 500 600 ... would indicate 100 doses of vaccine 1 to become available of day 1 or the response, 300 on day 2, etc. for up to the number of days specified in run length.

antiviral doses daily: integer
Number of antiviral courses that can be delivered daily by available resources. Variable

vaccine doses daily: integer
Number of vaccinations that can be administered daily by available resources. Variable

vaccine data: integer, real[3], real[6], bool
Total of 11 values indicating the vaccine id, vaccine efficacy and administration policies for each vaccine. The vaccine efficacy parameters (VE) should be between 0 (no efficacy) and 1 (complete efficacy). The administration policy values are the fraction of the age groups (infant, pre-school age, school-age, adult, and elderly) and pregnant women who are restricted from getting the vaccine. The simulation randomly assigns this fraction of the individuals in these groups to be restricted. The default is 0 (none restricted), while 1 indicates that 100% of individuals in this class are restricted. Values must be entered in the following order: integer, Numeric ID for the vaccine, starting from 0,
real, VE_S (the vaccine efficacy for susceptibility),
real, VE_I (the vaccine efficacy for infectiousness),
real, VE_P (the vaccine efficacy for illness given infection),
real, Fraction of infants (age < 6 months) restricted from getting the vaccine,
real, Fraction of pre-school age children (ages 0-4) restricted from getting the vaccine,
real, Fraction of school age children (ages 5-18) restricted from getting the vaccine,
real, Fraction of young adults (ages 19-29) restricted from getting the vaccine,
real, Fraction of older adults (ages 30-64) restricted from getting
the vaccine,
real, Fraction of elderly (ages 65+) restricted from getting the vaccine,
bool, Pregnant adults restricted from getting the vaccine.

An example of specifying multiple vaccines is vaccinedata 0 0.4 0.5 0.83 1 0.2 0.2 1 1 1, or vaccinedata 1 0.4 0.4 0.67 0 0 0 0 0 0.

Here vaccine 0 could be a live vaccine for children and vaccine 1 could be administered to anyone. Note that vaccine 0 has 0.2 for the child restrictions, which means that 20% of children (chosen at random) are not eligible.

vaccinebuildup: integer, integer, real[29]

Total of 31 values. The first value is the vaccine numeric ID, and the second is the day that the boost should be given. The remaining 29 values are the vaccine efficacy over the 29 days after the vaccine is given.

integer, Numeric ID for the vaccine. First vaccine must start at 0.
integer, Minimum number of days between the prime and boost ranging from 0 to 28. Default is 0, no boost.
reals, 29 values describing the vaccine efficacy buildup, ranging from 0 to 1. The value on the last day should be 1. The default is a one-dose vaccine that reaches maximum efficacy in two weeks.

vaccineefficacybyage: real[5]

Vector of five values representing the relative vaccine efficacy for each age group. Age groups are, in order: pre-school (0-4 years), school-age children (5-18 years), young adults (19-29 years), older adults (30-64 years), and elderly (65+ years). Values can range from 0 = no efficacy to 1 = full efficacy. The default is 1, full efficacy, for all age groups. For example, "1 1 1 1 0.6" would make vaccines only 60% as effective in the elderly as everyone else. The same settings are used for all vaccines.

AVEs: real
Antiviral vaccine efficacy for susceptibility (VE_S).

AVEi: real
Antiviral vaccine efficacy for infectiousness (VE_I).

AVEp: real
Antiviral vaccine efficacy for illness given infection (VE_P).

responsethreshold: real
Fraction of the population ascertained that results in initiating reactive strategies. Reactive strategies include vaccinations, deploying antiviral, and non-pharmaceutical interventions like liberal leave. For example, 0.01 would set this trigger at 1% of the population. The default is 0.0, which initiates reactive strategies after the first person is ascertained.

responsedelay: integer
Number of days to wait before initiating reactive strategies. A Decision Variable value of -1 would deploy reactive strategies on day 0, the first day of the simulation. This differs from the pre-strategy option because pre-vaccination assumes that people can be vaccinated early.
enough such that they have full protection on day 0. With a
response delay of -1, they might get vaccine on day 0.

ascertainment delay: integer Number of days it takes medical personnel to ascertain a symptomatic individual.

ascertainment fraction: real Fraction of all symptomatic individuals who will be ascertained. 0.7

essential fraction: real Fraction of working-age adults that belong to the "essential workforce" and are prioritized for receiving vaccine. The essential workforce is 6.9% of the employed population when there is a vaccine shortage and 10.8% if there is not a shortage. Default is 0, no essential workers prioritized to get vaccine.

pregnant fraction: real[5] Fraction of people in each of the 5 age groups who are pregnant. 0 0 0.02771 0.02069 0

high risk fraction: real[5] Fraction of individuals in each of the 5 age groups who are at high-risk of complications from influenza. For example, 0.089 0.089 0.212 0.212 0.0 would make 8.9% of children and 21.2% of 0.0 adults under 65 years high-risk.

seed tract: integer[4] Seeds a single census tract with infected people. State, county, and tract FIPS followed by the number of people to infect.

seed infected: integer Number of people to infect across the whole population. 12

seed infected daily: binary Indicates if infected people should be introduced into the population every day or just on the first day. A value of 1 will seed infected people every day, while 0 seeds infected only on day 0.

seed airports: integer Value between 0 and 10000 to indicate the number of passengers per 10000 per day to infect in airports.

travel: binary Enables short-term long-distance travel. Value of 1 enables travel, 0 indicates no travel. Default is 0, but should be set to 1 if simulating the continental US.

antiviral policy: string Indicates which people get antivirals. Can be one of four possible HHTAP values: none, treatment only, HHTAP (household members all get drugs if one member is ascertained), or HHTAP100 (special option: drugs can go to first 100 households that have a member ascertained).

school closure policy: string Identify which schools to close. Can take one of three possible none values, none, all, or by tract and age. An epidemic is detected after the response threshold and response delay variable criteria are met. At this point, either no schools close (none), all schools close (all) or the schools in a single tract that correspond to a single age group (e.g., elementary, middle, or high) can be closed (by tract and age).
schoolclosedays: integer Number of days to close schools ranging from 0 to the value of runlength. The default is 0, no school closure. A value larger than runlength would close the schools permanently.

isolation: real Voluntary isolation compliance probability. This is the probability that a sick person stays home voluntarily.

liberalleave: real Liberal leave compliance probability. The probability that people will take off from work if they get sick.

quarantine: real Voluntary household quarantine compliance probability. This is the probability that the members of a household stay home if one of the family members gets sick.

schoolopening: integer[56] School opening day for each state after the start of the simulation. Values of 0 or -1 indicate that the state's schools are open when the simulation starts, while other values indicate that the state's schools' opening day. The parameter takes 56 arguments, which correspond to the FIPS codes of the states.

2 DALYs Function Calibration (dq₁ Sensitivity Analysis)

The DALYs function computes the total disability-adjusted life years lived by individuals in different health states over the course of the simulation. The disability-adjustment coefficients are defined such that the healthier individuals have lower disability (dq_{asymptomatic} ≤ dq_{symptomatic}). In order to calibrate these coefficients in our example, we assume the binding budget constraint of v₁ + v₂ = 40,000 (so that the whole budget is spent), and perform a one-way sensitivity analysis of model outputs to different values of dq₁. Figures 1 and 2 plot the values as a function of investment v₂, with a fixed value of dq₂= 2. As expected, AR is insensitive to dq₁. The AR values in this case are very close, and the plots are essentially on top of each other. On the other hand, dq₂ affects the value of DALYs; with higher dq₂ the function is higher.
In summary, AR values are insensitive to the variation of $dq_1$ and $dq_2$ coefficients, and minor fluctuations are due to simulation-output randomness. On the other hand, the DALYs values are sensitive to the coefficients’ values, with $dq_1$ affecting DALYs values, and $dq_2$ affecting the function’s values and shape. In both cases, however, the horizontal location of the DALYs function is fixed, and changes in the function's convexity do not affect the minimization valley. Thus, we don't expect the $dq_i$ coefficients to play a crucial role in approximating the optimal solution of the RA problem. Conventionally, we choose $dq_1 = 1$ and $dq_2 = 2$, which agrees with our prior influenza model's assumption in Section 4.1 of
Kasaie and Kelton (2013) (symptomatic individuals are twice as infectious as are asymptomatic people).

3 Variance Reduction

The main drawback of using stochastic simulation models is the uncertainty of the results and the associated variance of the simulation outputs. In order to save computation cost and time, we try the common random numbers (CRN) variance-reduction technique to reduce the required number of runs of our ABS epidemic model. CRN requires synchronization of the random-number streams, which ensures that, in addition to using the same random numbers to simulate all configurations, a specific random number that is used for a specific purpose in one configuration is used for exactly the same purpose in all other configurations.

Following the suggested guidelines in Law and Kelton (2000), we identify 46 instances of random inputs in the ABS model of influenza (FLuTE) introduced in Section 4 of Kasaie and Kelton (2013). These instances correspond to different sources of randomness in the model, such as creating the original population, randomizing the age groups, setting the individual characteristics of a heterogeneous population (e.g., initial disease state, job status, disease resistance), and modeling the disease transmission through random contacts.

3.1 Random-Number Generator (RNG)

The originally available copy of the model uses a double precision SIMD-oriented Fast Mersenne Twister (dSFMT) based on the IEEE 754 format. We replace that current RNG with a combined multiple-recursive generator developed by L’Ecuyer et al. (2001). This generator supports multiple generators (streams) running simultaneously, while each generator (stream) has its sequence of numbers partitioned into many long disjoint
contiguous sub-streams, with no practical possibility of the streams’ ever overlapping each other given their enormous length. The basic underlying generator for this implementation is a combined multiple-recursive generator with period length of approximately $2^{191} \approx 3.1 \times 10^{57}$, proposed by L’Ecuyer (1999). We use a publicly available C++ interface of the code and apply it to the simulation model.

3.2 Testing CRN Efficiency

In order to study the efficiency of the CRN synchronization technique, we then consider two main strategies with respect to the random structure of the model.

1. Using independent sampling and reading independent random streams to each random instance in each model run. This will ensure the complete randomness of the input variates and the independence of the outputs.

2. Using CRN and reading the same random stream to each random instance through different model runs. This will result in correlated outputs that are expected to have a lower overall variance in their difference.

The implementation steps for each strategy are as follows:

- **Strategy 1: Independent Sampling**
  1. Initialize 46 independent random streams $RS = \{RS_1, \ldots, RS_{46}\}$, where $RS_i$ is the $i^{th}$ random stream.
  2. Reset the all the random streams to an initial seed vector $Z_0 = \{z_1, \ldots, z_{46}\}$, where $z_i$ is the random initial seed for $RS_i, i=1,\ldots,46$.
  3. For each simulation run, reset the all the streams to the next sub-stream.

Figure 3 shows the independent-sampling logic and the implementation procedure for this scenario where the $d_j, j=1,\ldots,n$ is the $j^{th}$ experimental design point, and $r$ is the number of replications for each point.
Strategy 2: CRN

1. Initialize 46 independent random streams $RS = \{RS_1, \ldots, RS_{46}\}$.

2. Reset the random streams to an initial seed $Z_0 = \{z_1, \ldots, z_{46}\}$, where $z_i$ is the random initial seed for $RS_i$, $i = 1, \ldots, 46$.

3. For each replication, reset all the streams to the next sub-stream.

4. For each experimental point, reset the all random streams to the initial seed $Z_0$.

In order to account for different length of random streams used by each simulation run, the streams are reset to the next sub-stream at the beginning of each replication. Figure 4 shows the synchronization logic in this approach for the first two design points.
In this setting, the random-number streams are reset to $Z_0$ for each design point, so that the first replication of each design point reads the same random numbers from the first sub-stream. The second replication is then initialized by resetting all of the streams to the next sub-streams within each stream, and same procedure is used for all future replications. Therefore, regardless of the models’ structure and number of random variates used, the $i^{th}$, $i$
= 1, ..., \( r \), replication of each design point starts running from the same \( i^{th} \) sub-stream and uses the same random variates.

### 3.3 CRN Experiment

We randomly select seven design points \((d_i, i = 1, \ldots, 7)\) in the models’ parameter space and take 14 pairs of these points randomly, \((d_{ij} = d_i - d_j, i \neq j, i, j = 1, \ldots, 7, k = 1, \ldots, 14)\). The two strategies (independent sampling and CRN) are tried for each design point, and the pairs’ outputs’ differences (\(AR_k = AR_i - AR_j \) and \(DALY_{sk} = DALY_{si} - DALY_{sj}\)) are computed under each scenario. We use \( r = 400 \) replications for our simulation. The output columns corresponding to the same pair’s output difference under the two strategies (e.g., \(AR_k\) under independent sampling vs. CRN) are then compared with each other, and an \( F \)-test is used to test for a significant reduction in the variance of the CRN case. Figures 5 and 6 show the comparison of the means and standard deviations of the pairs’ output differences under each scenario.
Table 2 shows the $p$-values of the one-tailed $F$-test for testing the equality of variances of outputs’ differences for each pair under the two scenarios. The $F$-critical value under the 0.05 level of significance is $F_{0.05, 399, 399} = 1.1792$.

Figure 5: Comparison of Means of Pairs’ Output Differences under Each Scenario.

Figure 6: Comparison of Standard deviation of Pairs’ Output Differences under Each Scenario.
The graphical comparison of the output differences shows a significant reduction of standard deviations under the synchronized case, and the one-tailed $F$-test for equality of variances rejects the null hypothesis in all cases. The CRN technique results in an average relative reduction of 72% in the standard deviation of outputs.

The significant reduction in the variability of differences in outputs using CRN is a result of induced positive correlation among design points. The CRN strategy however, does not affect the variances of the individual outputs, so that the required number of runs to achieve a specific relative precision for each design point itself is similar under both strategies. Moreover, the induced positive correlation violates the independence assumption made in the analysis of experimental designs and the response-surface methodology (RSM) technique. With regard to these issues and the satisfactory speed of the simulation model under the current settings (for a small population of 2000 people) to achieve acceptable

Table 2: One-Tailed F-Test’s P-Values.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$P(\text{F} \leq f)$ one-tail AR</th>
<th>DALYs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.83E–105</td>
<td>5.35E–121</td>
</tr>
<tr>
<td>2</td>
<td>1.99E–133</td>
<td>2.94E–136</td>
</tr>
<tr>
<td>3</td>
<td>3.18E–177</td>
<td>6.83E–143</td>
</tr>
<tr>
<td>4</td>
<td>1.38E–183</td>
<td>6.18E–130</td>
</tr>
<tr>
<td>5</td>
<td>4.11E–187</td>
<td>1.17E–127</td>
</tr>
<tr>
<td>6</td>
<td>1.14E–193</td>
<td>8.80E–135</td>
</tr>
<tr>
<td>7</td>
<td>5.21E–53</td>
<td>2.18E–76</td>
</tr>
<tr>
<td>8</td>
<td>9.89E–82</td>
<td>7.29E–96</td>
</tr>
<tr>
<td>9</td>
<td>1.64E–100</td>
<td>1.34E–105</td>
</tr>
<tr>
<td>10</td>
<td>7.04E–131</td>
<td>3.82E–117</td>
</tr>
<tr>
<td>11</td>
<td>1.62E–158</td>
<td>2.51E–127</td>
</tr>
<tr>
<td>12</td>
<td>8.46E–99</td>
<td>5.57E–112</td>
</tr>
<tr>
<td>13</td>
<td>2.44E–63</td>
<td>1.28E–80</td>
</tr>
<tr>
<td>14</td>
<td>4.33E–119</td>
<td>1.64E–121</td>
</tr>
</tbody>
</table>
precision, we prefer to adopt the simpler and statistically “cleaner” independent-sampling approach for our further experimentations.

4 Reading the Triangular Coordinates in Mixture Designs

The investment values in the influenza pandemic RA problem are subject to a global budget constraint and are individually bounded to certain ranges (Section 4.4.1 in Kasaie and Kelton, 2013). We use a D-optimal mixture design to generate the experiments, and we summarize the results via triangular (or ternary) graphs in each case. In this section, we provide a brief description of these graphs to assist understanding and reading-off of the point’s coordinates, and refer the reader to Hill and Lewicki (2006) for more information and further examples.

Suppose a mixture consists of 3 components A, B, and C. Any mixture of the three components can be summarized by a point in the triangular coordinate system defined by the three variables. The values for the components in each mixture can be interpreted as proportions so that the sum for each mixture is 1.0. If we graph these data in a regular 3D scatter plot, it becomes apparent that the points form a triangle in the 3D space. Only the points inside the triangle where the sum of the component values is equal to 1 are valid mixtures. Therefore, we can simply plot only the triangle to summarize the component values (proportions) for each mixture.

To read off the coordinates of a point in the triangular graph, we drop a line from each respective vertex to the opposite side of the triangle. At the vertex for the particular factor, there is a pure blend, that is, one that contains only the respective component. Thus, the coordinates for the vertex point is 1 (or 100%, or however else the mixtures are scaled) for the respective component, and 0 for all other components. At the side opposite to the
respective vertex, the value for the respective component is 0 (zero), and 0.5 (or 50%, etc.) for the other components as shown in Figure 7.

In our example, the investment amounts in each intervention, \( v_i \), \( i = 1, \ldots, 6 \), are subject to a global budget constraint and can be considered as component of a mixture design. Figure 8 shows the desirability plot for the mixture design presented in Section 4.4.1 of Kasaie and Kelton (2013). The three factors \( (v_1, v_2, v_3) \) are bounded in a triangular lattice formed by the binding budget constraint so that the sum of investment ratios equals 0.671, with \( v_4 = 0.329 \) and \( v_5 = v_6 = 0 \). The shades represent the desirability values across the feasible area of the lattice, and the near-optimal investment is numerically estimated at \((0.49, 0.14, 0.04, 0.329, 0.00, 0.00)\) with a desirability of 0.951. The coordinates of this near-optimal point can be read by drawing the perpendicular lines to the triangle’s altitudes, and reading the corresponding investment proportions.

Figure 7: The Triangular Mixture Graph.
5 Numerical Results

In this section, we provide the results of our analysis for the numerical example discussed in Section 4 of Kasaie and Kelton (2013). We use experimental designs to study the behavior of the model’s objective functions, $R1s$ and $R2s$ corresponding to the standardized value of AR and DALYs functions, and follow the RSM procedure to attempt to optimize the simulation outputs.

RSM is a sequential procedure that explores small (local) sub-regions of the experimental region and follows the direction of improvement. At each iteration, an experimental design is selected, and local meta-models are fit to each response. These models are then used to estimate the direction of improvement. In our example, we follow the steepest-descent direction for $R2s$ values while controlling for the behavior of $R1s$. The search is then continued in this direction until no further improvement in the responses is observed. At this point, a new iteration begins and new meta-models are fit to the local sub-region.
Although the common RSM approach does not guarantee finding the true global optimum, a quick remedy is using a multi-start approach to increase reliability of the suggested solution (Montgomery, 2001). We therefore repeat our analysis for three different initial points selected from the feasible space of the RA problem. The results show that the RSM procedure converges to the same investment strategy in all the three experiments, \( \hat{\rho} \approx (0.63, 0.00, 0.20, 0.17, 0.00, 0.00) \) so that we adopt the suggested point as the near-optimal solution of the RA problem.

In the following we provide the numerical results of our analysis for each experiment. The following analyses are performed using the Design-Expert statistical software (StatEase, 2010), and the significance level is assumed to be 0.05.

### 5.1 Experiment 1

We select the best investment strategy suggested by mixture design desirability function (Section 4.4.1), as the initial point of RSM, and start our search from this point, \( \hat{\rho} = (0.487, 0.142, 0.420, 0.328, 0.000, 0.000) \).

Tables 3 through 8 summarize the results for each iteration of the RSM procedure (fitting the appropriate polynomial and following the SD path to improve the outputs’ performances). Figures 9 and 10 show the contour plots of the nonlinear fitted models on the outputs in iteration 3 and 4 respectively. Figure 11 plots the desirability function of the model’s outputs in the final iteration, which is maximized at the near-optimal suggested point \( \hat{\nu} = (25276.723, 0.000, 8000.000, 6723.277, 0.000, 0.000) \), corresponding to \( \hat{\rho} = (0.632, 0.000, 0.200, 0.168, 0.000, 0.000) \).
The binding budget constraint reduces the problem dimension by one, so that one binding variable is defined as a function of the total budget and the other independent variables (e.g., $v_4 = 40000 - (v_1 + v_2 + v_3 + v_5 + v_6)$).
### Table 5: Experiment 1- Iteration 2.

<table>
<thead>
<tr>
<th>Design Information:</th>
<th>Design Type</th>
<th># of Design Points</th>
<th># of Infeasible Points</th>
<th># of C.P.s</th>
<th># of Replications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Half CCD</td>
<td>32</td>
<td>5</td>
<td>6</td>
<td>500</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design Points:</th>
<th>v₁</th>
<th>v₂</th>
<th>v₃</th>
<th>v₄</th>
<th>v₅</th>
<th>v₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Point</td>
<td>21091.13</td>
<td>5000.00</td>
<td>8500.00</td>
<td>5408.87</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Upper Limit</td>
<td>15000.00</td>
<td>0.00</td>
<td>8000.00</td>
<td>binding</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Lower Limit</td>
<td>25000.00</td>
<td>10000.00</td>
<td>10000.00</td>
<td>binding</td>
<td>2000.00</td>
<td>2000.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fit Results:</th>
<th>R₁s</th>
<th>R₂s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Rsqr</td>
<td>Rsqr-adj</td>
</tr>
<tr>
<td>Linear</td>
<td>0.28</td>
<td>0.14</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R₁s</th>
<th>Source</th>
<th>P-values</th>
<th>R₂s</th>
<th>Source</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.60610</td>
<td>v₁</td>
<td>0.1083</td>
<td>0.88235</td>
<td>v₁</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>−0.03217</td>
<td>v₂</td>
<td>0.0177</td>
<td>−0.03393</td>
<td>v₂</td>
<td>0.0355</td>
</tr>
<tr>
<td>−0.02246</td>
<td>v₃</td>
<td>0.0889</td>
<td>−0.00819</td>
<td>v₃</td>
<td>0.7669</td>
</tr>
<tr>
<td>0.00773</td>
<td>v₄</td>
<td>0.5484</td>
<td>0.00553</td>
<td>v₄</td>
<td>0.9141</td>
</tr>
<tr>
<td>−0.00099</td>
<td>v₅</td>
<td>0.9383</td>
<td>0.00201</td>
<td>v₅</td>
<td>0.8726</td>
</tr>
<tr>
<td>−0.00580</td>
<td>v₆</td>
<td>0.6519</td>
<td>−0.00299</td>
<td>v₆</td>
<td></td>
</tr>
</tbody>
</table>

### Table 6: Experiment 1- Iteration 2-5D Summary.

<table>
<thead>
<tr>
<th>Leading Response</th>
<th>Leading Factor</th>
<th>Step Size</th>
<th>Replications</th>
<th># of steps</th>
<th>Improvement ceases at step</th>
</tr>
</thead>
<tbody>
<tr>
<td>R₁s</td>
<td>v₁</td>
<td>500</td>
<td>500</td>
<td>12</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Suggested Point</th>
<th>v₁</th>
<th>v₂</th>
<th>v₃</th>
<th>v₄</th>
<th>v₅</th>
<th>v₆</th>
<th>R₁s</th>
<th>R₂s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25091.12648</td>
<td>0.00000</td>
<td>8000.00000</td>
<td>6908.87352</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.54491</td>
<td>0.07972</td>
</tr>
</tbody>
</table>
Table 7: Experiment 1 - Iteration 3.

<table>
<thead>
<tr>
<th>Design Information:</th>
<th>Design Type</th>
<th># of Design Points</th>
<th># of Infeasible Points</th>
<th># of C.P.s</th>
<th># of Replications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D-Optimal</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>1000</td>
</tr>
<tr>
<td></td>
<td>Constrained</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design Points:</th>
<th>v₁</th>
<th>v₂</th>
<th>v₃</th>
<th>v₄</th>
<th>v₅</th>
<th>v₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Point</td>
<td>25091.13</td>
<td>0.00</td>
<td>8000.00</td>
<td>6908.87</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Upper Limit</td>
<td>24500.00</td>
<td>0.00</td>
<td>7500.00</td>
<td>6500.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Lower Limit</td>
<td>25500.00</td>
<td>0.00</td>
<td>8500.00</td>
<td>7500.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fit Results:</th>
<th>R₁s</th>
<th>R₂s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rsqr</td>
<td>Rsqr-adj</td>
</tr>
<tr>
<td>Model</td>
<td>Quadratic</td>
<td>Linear</td>
</tr>
<tr>
<td>Rsqr</td>
<td>0.94</td>
<td>0.49</td>
</tr>
<tr>
<td>Rsqr-adj</td>
<td>0.79</td>
<td>0.38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R₁s Source P-values</th>
<th>R₂s Source P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00000 Model</td>
<td>0.00014 Model</td>
</tr>
<tr>
<td>-0.00009 Linear</td>
<td>-0.00000 Linear</td>
</tr>
<tr>
<td>-0.00048 v₃</td>
<td>-0.00001 v₃</td>
</tr>
<tr>
<td>-0.00099 v₄</td>
<td>0.00001 v₄</td>
</tr>
<tr>
<td>0.00000 v₁*v₃</td>
<td>0.00003 v₁*v₃</td>
</tr>
<tr>
<td>0.00000 v₁*v₄</td>
<td>0.00010 v₁*v₄</td>
</tr>
<tr>
<td>0.00000 v₃ v₄</td>
<td>0.00010 v₃ v₄</td>
</tr>
</tbody>
</table>
Figure 9: R1s (Left) and R2s (Right) Fitted Model's Contours at Iteration 3.

Table 8: Experiment 1- Iteration 4 (Final Experiment).

<table>
<thead>
<tr>
<th>Design Information:</th>
<th>Design Type</th>
<th># of Design Points</th>
<th># of Infeasible Points</th>
<th># of C.P.s</th>
<th># of Replications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Two factor Mixture</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design Points:</th>
<th>v_1</th>
<th>v_2</th>
<th>v_3</th>
<th>v_4</th>
<th>v_5</th>
<th>v_6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Point</td>
<td>24554.45</td>
<td>0.00</td>
<td>8363.25</td>
<td>7082.30</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Upper Limit</td>
<td>24500.00</td>
<td>0.00</td>
<td>8000.00</td>
<td>6500.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Lower Limit</td>
<td>25500.00</td>
<td>0.00</td>
<td>8000.00</td>
<td>7500.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Fit Results:

<table>
<thead>
<tr>
<th></th>
<th>R1s</th>
<th>R2s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Rsqr</td>
<td>Rsqr-adj</td>
</tr>
<tr>
<td>Cubic</td>
<td>0.66</td>
<td>0.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R1s Source</th>
<th>P-values</th>
<th>R2s Source</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>0.1929</td>
<td>Model</td>
<td>0.00000</td>
</tr>
<tr>
<td>Linear Mixture</td>
<td>0.6928</td>
<td>Linear Mixture</td>
<td>0.0005</td>
</tr>
<tr>
<td>v_1</td>
<td>–</td>
<td>v_1</td>
<td>–</td>
</tr>
<tr>
<td>v_4</td>
<td>–</td>
<td>v_4</td>
<td>–</td>
</tr>
<tr>
<td>v_1*v_4</td>
<td>0.7938</td>
<td>v_1*v_4</td>
<td>0.0081</td>
</tr>
<tr>
<td>v_1*(v_1 - v_4)</td>
<td>0.0527</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.2 Experiment 2

We select the best design point in the mixture design as the initial point for the RSM, and start our search from this point, $\tilde{v} = (29128.15, 842.28, 52.66, 9294.42, 0.00, 682.49)$.
Tables 9 through 13 summarize the results for each iteration of the RSM procedure. Figure 12 shows the contour plots of the nonlinear fitted models on the outputs in iteration 3, and Figure 13 plots the desirability function of the model’s outputs in the final iteration. The final near-optimum is estimated at $\vec{v} = (25000.00, 0.00, 8310.79, 6689.21, 0.00, 0.00)$, corresponding to $\vec{\rho} = (0.62, 0.00, 0.21, 0.17, 0.00, 0.00)$.

Table 9: Experiment 2- Iteration 1.

<table>
<thead>
<tr>
<th>Design Information:</th>
<th>Design Type</th>
<th># of Design Points</th>
<th># of Infeasible Points</th>
<th># of C.P.s</th>
<th># of Replications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two Level</td>
<td>Factorial</td>
<td>32</td>
<td>4</td>
<td>6</td>
<td>400</td>
</tr>
</tbody>
</table>

Design Points:

- Initial Point: $v_1 = 29128.15$, $v_2 = 842.28$, $v_3 = 52.66$, $v_4 = 9294.42$, $v_5 = 0.00$, $v_6 = 682.49$
- Upper Limit: $v_1 = 31000.00$, $v_2 = 5000.00$, $v_3 = 2000.00$, $v_4 = binding$, $v_5 = 2000.00$, $v_6 = 2000.00$
- Lower Limit: $v_1 = 27000.00$, $v_2 = 0.00$, $v_3 = 0.00$, $v_4 = binding$, $v_5 = 0.00$, $v_6 = 0.00$

Fit Results:

- $R1s$:
  - Linear: $Rsqr = 0.99$, $Rsqr-adj = 0.99$
  - $R2s$:
    - Linear: $Rsqr = 0.99$, $Rsqr-adj = 0.99$

<table>
<thead>
<tr>
<th>Source</th>
<th>$P$-values</th>
<th>Source</th>
<th>$P$-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$&lt; 0.0001$</td>
<td>Model</td>
<td>$0.16033$</td>
</tr>
<tr>
<td>$v_1$</td>
<td>$&lt; 0.0001$</td>
<td>$v_1$</td>
<td>$0.0223$</td>
</tr>
<tr>
<td>$v_2$</td>
<td>$&lt; 0.0001$</td>
<td>$v_2$</td>
<td>$0.1441$</td>
</tr>
<tr>
<td>$v_3$</td>
<td>$&lt; 0.0001$</td>
<td>$v_3$</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td>$v_4$</td>
<td>$&lt; 0.0001$</td>
<td>$v_4$</td>
<td>$0.2623$</td>
</tr>
<tr>
<td>$v_5$</td>
<td>$&lt; 0.0001$</td>
<td>$v_5$</td>
<td>$0.5562$</td>
</tr>
</tbody>
</table>
### Table 10: Experiment 2 - Iteration 1 - SD Summary.

<table>
<thead>
<tr>
<th>Leading Response</th>
<th>Leading Factor</th>
<th>Step Size</th>
<th>Replications</th>
<th># of steps</th>
<th>Improvement ceases at step</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R^2_s )</td>
<td>( v_3 )</td>
<td>1000</td>
<td>1000</td>
<td>10</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( v_1 )</th>
<th>( v_2 )</th>
<th>( v_3 )</th>
<th>( v_4 )</th>
<th>( v_5 )</th>
<th>( v_6 )</th>
<th>( R^1_s )</th>
<th>( R^2_s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suggested Point</td>
<td>28407.45</td>
<td>0.00</td>
<td>9000.00</td>
<td>2592.55</td>
<td>0.00</td>
<td>0.00</td>
<td>0.58612</td>
<td>0.080</td>
</tr>
</tbody>
</table>

### Table 11: Experiment 2 - Iteration 2.

<table>
<thead>
<tr>
<th>Design Information:</th>
<th>Design Type</th>
<th># of Design Points</th>
<th># of Infeasible Points</th>
<th># of C.P.s</th>
<th># of Replications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half CCD</td>
<td>13</td>
<td>0</td>
<td>5</td>
<td>1000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design Points:</th>
<th>( v_1 )</th>
<th>( v_2 )</th>
<th>( v_3 )</th>
<th>( v_4 )</th>
<th>( v_5 )</th>
<th>( v_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Point</td>
<td>28407.45</td>
<td>0.00</td>
<td>9000.00</td>
<td>2592.55</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Upper Limit</td>
<td>27000.00</td>
<td>0.00</td>
<td>8000.00</td>
<td>binding</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Lower Limit</td>
<td>29000.00</td>
<td>0.00</td>
<td>10000.00</td>
<td>binding</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fit Results:</th>
<th>( R^1_s )</th>
<th>( R^2_s )</th>
<th>( R^1_s )</th>
<th>( R^2_s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Rsqr</td>
<td>Rsqr-adj</td>
<td>Model</td>
<td>Rsqr</td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.95</td>
<td>0.91</td>
<td>Quadratic</td>
<td>0.65</td>
</tr>
</tbody>
</table>

\[
\begin{array}{cccc}
\text{\( R^1_s \)} & \text{Source} & \text{P-value} & \text{\( R^2_s \)} & \text{Source} & \text{P-value} \\
-0.80073 & \text{Model} & 0.0002 & -0.54059 & \text{Model} & 0.1196 \\
0.11172 & \text{\( v_1 \)} & < 0.0001 & 0.06393 & \text{\( v_1 \)} & 0.4072 \\
-0.08870 & \text{\( v_3 \)} & 0.0011 & -0.06207 & \text{\( v_3 \)} & 0.1428 \\
-0.00108 & \text{\( v_1 \) * \( v_3 \)} & 0.6303 & 0.00008 & \text{\( v_1 \) * \( v_3 \)} & 0.9583 \\
-0.00159 & \text{\( v_1 \) }^2 & 0.3610 & -0.00114 & \text{\( v_1 \) }^2 & 0.3567 \\
0.00705 & \text{\( v_3 \) }^2 & 0.0034 & 0.00322 & \text{\( v_3 \) }^2 & 0.0268 \\
\end{array}
\]
### Table 12: Experiment 2 - Iteration 2 - SD Summary.

<table>
<thead>
<tr>
<th>Leading Response</th>
<th>Leading Factor</th>
<th>Step Size</th>
<th>Replications</th>
<th># of steps</th>
<th>Improvement ceases at step</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_1$s</td>
<td>$v_1$</td>
<td>500</td>
<td>1000</td>
<td>6</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$v_1$</th>
<th>$v_2$</th>
<th>$v_3$</th>
<th>$v_4$</th>
<th>$v_5$</th>
<th>$v_6$</th>
<th>$R_1$s</th>
<th>$R_2$s</th>
</tr>
</thead>
<tbody>
<tr>
<td>25500.00</td>
<td>0.00</td>
<td>8246.86</td>
<td>6253.14</td>
<td>0.00</td>
<td>0.00</td>
<td>0.55881</td>
<td>0.07929</td>
</tr>
</tbody>
</table>

### Table 13: Experiment 2 - Iteration 3 (Final Design).

<table>
<thead>
<tr>
<th>Design Information:</th>
<th>Design Type</th>
<th># of Design Points</th>
<th># of Infeasible Points</th>
<th># of C.P.s</th>
<th># of Replications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Half CCD</td>
<td>13</td>
<td>0</td>
<td>5</td>
<td>1000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design Points:</th>
<th>$v_1$</th>
<th>$v_2$</th>
<th>$v_3$</th>
<th>$v_4$</th>
<th>$v_5$</th>
<th>$v_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Point</td>
<td>25500.00</td>
<td>0.00</td>
<td>8246.86</td>
<td>6253.14</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Upper Limit</td>
<td>25000.00</td>
<td>0.00</td>
<td>7500.00</td>
<td><em>binding</em></td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Lower Limit</td>
<td>26000.00</td>
<td>0.00</td>
<td>8500.00</td>
<td><em>binding</em></td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fit Results:</th>
<th>$R_1$s</th>
<th>$R_2$s</th>
<th></th>
<th>$R_1$s</th>
<th>$R_2$s</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Rsqr</td>
<td>Rsqr-adj</td>
<td></td>
<td>Model</td>
<td>Rsqr</td>
<td>Rsqr-adj</td>
</tr>
<tr>
<td>Quadratic</td>
<td>0.71</td>
<td>0.49</td>
<td></td>
<td>Quadratic</td>
<td>0.89</td>
<td>0.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$R_1$s Source</th>
<th>P-value</th>
<th>$R_2$s Source</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>0.07230</td>
<td>Model</td>
<td>0.0028</td>
</tr>
<tr>
<td>$v_1$</td>
<td>0.4725</td>
<td>$v_1$</td>
<td>0.6167</td>
</tr>
<tr>
<td>$v_3$</td>
<td>0.0208</td>
<td>$v_3$</td>
<td>0.0004</td>
</tr>
<tr>
<td>$v_1$v_3</td>
<td>0.6767</td>
<td>$v_1$v_3</td>
<td>0.6844</td>
</tr>
<tr>
<td>$v_1^2$</td>
<td>0.6812</td>
<td>$v_1^2$</td>
<td>0.5562</td>
</tr>
<tr>
<td>$v_3^2$</td>
<td>0.0309</td>
<td>$v_3^2$</td>
<td>0.0048</td>
</tr>
</tbody>
</table>
Experiment 3

In this experiment, we choose a random initial point from the RA feasible space and modified it to the nearest investment point so that it covers at least one level of each intervention’s initial investment requirement, $v = (6000, 7000, 4000, 4000, 3500, 3500)$. 

5.3 Experiment 3

Figure 12: R1s (Left) and R2s (Right) Fitted Model's Contours at Iteration 3.

Figure 13: Desirability Function for R1s and R2s with Two Factors.
Tables 14 through 18 summarize the results for each iteration of the RSM procedure. Figure 14 plots the desirability function of the model’s outputs in the final iteration. The final optimum is estimated at $\hat{\theta} = (24000, 0, 8000, 8000, 0, 0)$, corresponding to $\hat{\rho} = (0.6, 0.0, 0.2, 0.2, 0.0, 0.0)$.

**Table 14: Experiment 3- Iteration 1.**

<table>
<thead>
<tr>
<th>Design Information:</th>
<th>Design Type</th>
<th># of Design Points</th>
<th># of Infeasible Points</th>
<th># of C.P.s</th>
<th># of Replications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half CCD</td>
<td>52</td>
<td>0</td>
<td>6</td>
<td>500</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design Points:</th>
<th>$v_1$</th>
<th>$v_2$</th>
<th>$v_3$</th>
<th>$v_4$</th>
<th>$v_5$</th>
<th>$v_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Point</td>
<td>6000.00</td>
<td>7000.00</td>
<td>4000.00</td>
<td>4000.00</td>
<td>3500.00</td>
<td>3500.00</td>
</tr>
<tr>
<td>Upper Limit</td>
<td>4000.00</td>
<td>4000.00</td>
<td>2000.00</td>
<td>2000.00</td>
<td>2000.00</td>
<td>2000.00</td>
</tr>
<tr>
<td>Lower Limit</td>
<td>8000.00</td>
<td>10000.00</td>
<td>6000.00</td>
<td>6000.00</td>
<td>5000.00</td>
<td>5000.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fit Results:</th>
<th>$R_1$s</th>
<th>$R_2$s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Rsqr</td>
<td>Rsqr</td>
</tr>
<tr>
<td>Linear</td>
<td>0.97</td>
<td>0.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$R_1$s</th>
<th>Source</th>
<th>P-value</th>
<th>$R_2$s</th>
<th>Source</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.05941</td>
<td>Model</td>
<td>&lt; 0.0001</td>
<td>1.17788</td>
<td>Model</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>-0.02269</td>
<td>$v_1$</td>
<td>&lt; 0.0001</td>
<td>-0.04494</td>
<td>$v_1$</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>-0.00058</td>
<td>$v_2$</td>
<td>0.2071</td>
<td>-0.00042</td>
<td>$v_2$</td>
<td>0.6218</td>
</tr>
<tr>
<td>-0.00024</td>
<td>$v_3$</td>
<td>0.7276</td>
<td>-0.00014</td>
<td>$v_3$</td>
<td>0.9104</td>
</tr>
<tr>
<td>-0.00809</td>
<td>$v_4$</td>
<td>&lt; 0.0001</td>
<td>-0.01360</td>
<td>$v_4$</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>-0.00036</td>
<td>$v_5$</td>
<td>0.6932</td>
<td>-0.00039</td>
<td>$v_5$</td>
<td>0.8188</td>
</tr>
<tr>
<td>-0.00060</td>
<td>$v_6$</td>
<td>0.5121</td>
<td>0.00006</td>
<td>$v_6$</td>
<td>0.9704</td>
</tr>
</tbody>
</table>

**Table 15: Experiment 3- Iteration 1- SD Summary.**

<table>
<thead>
<tr>
<th>Leading Response</th>
<th>Leading Factor</th>
<th>Step Size</th>
<th>Replications</th>
<th># of steps</th>
<th>Improvement ceases at step</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_2$s</td>
<td>$v_1$</td>
<td>2000</td>
<td>500</td>
<td>8</td>
<td>Later steps are infeasible</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$v_1$</th>
<th>$v_2$</th>
<th>$v_3$</th>
<th>$v_4$</th>
<th>$v_5$</th>
<th>$v_6$</th>
<th>$R_1$s</th>
<th>$R_2$s</th>
</tr>
</thead>
<tbody>
<tr>
<td>20000.00</td>
<td>4000.00</td>
<td>2000.00</td>
<td>8236.84</td>
<td>2000.00</td>
<td>2000.00</td>
<td>0.645310.26671</td>
<td></td>
</tr>
</tbody>
</table>
### Table 16: Experiment 3- Iteration 2.

<table>
<thead>
<tr>
<th>Design Information:</th>
<th>Design Type</th>
<th># of Design Points</th>
<th># of Infeasible Points</th>
<th># of C.P.s</th>
<th># of Replications</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Half CCD</td>
<td>32</td>
<td>4</td>
<td>6</td>
<td>1000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Design Points:</th>
<th>v₁ v₂ v₃ v₄ v₅ v₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Point</td>
<td>20000.00 4000.00 2000.00 8236.84 2000.00 2000.00</td>
</tr>
<tr>
<td>Upper Limit</td>
<td>18000.00 0.00 0.00 binding 0.00 0.00</td>
</tr>
<tr>
<td>Lower Limit</td>
<td>22000.00 5000.00 4000.00 binding 4000.00 4000.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fit Results:</th>
<th>RIₘ</th>
<th>RIₜ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model &amp; Rsqr</td>
<td>Linear</td>
<td>0.89</td>
</tr>
<tr>
<td>Rsqr-adj</td>
<td>Linear</td>
<td>0.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RIₘ</th>
<th>Source</th>
<th>P-value</th>
<th>RIₜ</th>
<th>Source</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.55790</td>
<td>Model</td>
<td>&lt; 0.0001</td>
<td>0.56251</td>
<td>Model</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>0.00404</td>
<td>v₁</td>
<td>0.0137</td>
<td>-0.01102</td>
<td>v₁</td>
<td>0.0137</td>
</tr>
<tr>
<td>0.00479</td>
<td>v₂</td>
<td>0.0016</td>
<td>0.00192</td>
<td>v₂</td>
<td>0.6011</td>
</tr>
<tr>
<td>-0.02008</td>
<td>v₃</td>
<td>&lt; 0.0001</td>
<td>-0.03634</td>
<td>v₃</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>0.00792</td>
<td>v₅</td>
<td>&lt; 0.0001</td>
<td>0.00456</td>
<td>v₅</td>
<td>0.3240</td>
</tr>
<tr>
<td>0.00440</td>
<td>v₆</td>
<td>0.0146</td>
<td>0.00190</td>
<td>v₆</td>
<td>0.6782</td>
</tr>
</tbody>
</table>

### Table 17: Experiment 3- Iteration 2 – SD Summary.

<table>
<thead>
<tr>
<th>Leading Response</th>
<th>Leading Factor</th>
<th>Step Size</th>
<th>Replications</th>
<th># of steps</th>
<th>Improvement ceases at step</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIₜ v₃</td>
<td>1000</td>
<td>500</td>
<td>12</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>RIₘ v₇ v₂ v₃ v₄ v₅ v₆</td>
<td>21818.89</td>
<td>0.00</td>
<td>8000.00</td>
<td>10181.11</td>
<td>0.00</td>
</tr>
</tbody>
</table>
### Table 18: Experiment 3- Iteration 3 (Final Design).

<table>
<thead>
<tr>
<th>Design Type</th>
<th># of Design Points</th>
<th># of Infeasible Points</th>
<th># of C.P.s</th>
<th># of Replications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half CCD</td>
<td>13</td>
<td>0</td>
<td>5</td>
<td>1000</td>
</tr>
</tbody>
</table>

**Design Points**

<table>
<thead>
<tr>
<th>Source</th>
<th>$v_1$</th>
<th>$v_2$</th>
<th>$v_3$</th>
<th>$v_4$</th>
<th>$v_5$</th>
<th>$v_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>21818.89</td>
<td>0</td>
<td>8000</td>
<td>10181.11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Upper</td>
<td>19000.00</td>
<td>0.00</td>
<td>6000.00</td>
<td>binding</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Lower</td>
<td>24000.00</td>
<td>0.00</td>
<td>8000.00</td>
<td>binding</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

**Fit Results:**

<table>
<thead>
<tr>
<th></th>
<th>$R1s$</th>
<th>$R2s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Rsqr</td>
<td>Rsqr-adj</td>
</tr>
<tr>
<td>2FI</td>
<td>0.57</td>
<td>0.42</td>
</tr>
<tr>
<td>Linear</td>
<td>0.99</td>
<td>0.98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$R1s$ Source</th>
<th>P-value</th>
<th>$R2s$ Source</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>0.0474</td>
<td>Model</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$v_1$</td>
<td>0.1140</td>
<td>$v_1$</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>$v_3$</td>
<td>0.0432</td>
<td>$v_3$</td>
<td>0.0310</td>
</tr>
<tr>
<td>$v_1$*$v_3$</td>
<td>0.1047</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 14:** Desirability Function for $R1s$ and $R2s$ with Two Factors.
6 References


---End of Chapter 1---
Chapter 2: Investigating the Dynamics of Tuberculosis Transmission and Effectiveness of Case-Finding Strategies Using an Agent-Based Simulation Approach
Essay 1: An Agent-Based Simulation of a Tuberculosis Epidemic: Understanding the Timing of Transmission

Parastu Kasaie¹, David. W. Dowdy², W. David Kelton¹,

¹ Department of Operations, Business Analytics, and Information Systems, College of Business, University of Cincinnati, Cincinnati, Ohio 45221, USA.
² Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21218, USA.

Abstract

Tuberculosis (TB) transmission is a key factor for disease-control policy, but the timing and distribution of transmission and the role of social contacts remain obscure. We develop an agent-based simulation of a TB epidemic in a single population, and consider a hierarchically structured contact network in three levels, typical of airborne diseases. The parameters are adopted from the literature, and the model is calibrated to a setting of high TB incidence. We model the dynamics of transmission at the individual level, and study the timing of secondary infections from a single source throughout the duration of the disease. We compare the patterns of transmission among different networks and discuss implications. Sensitivity analysis of outputs indicates the robustness of the results to variations in the parameter values.
1 Introduction

New and re-emerging infectious diseases pose an enormous burden on global health, and massive effort has been put into combating them. As an example, TB is a contagious bacterial infection responsible for 1.4 million deaths each year. About one third of the world’s population is infected with TB, and close to 9 million new TB cases occur each year (CORE Group 2006). Despite the vast literature and ongoing research on disease evolution, however, our understanding of the transmission dynamics is incomplete, and many questions remain unanswered (Dowdy, Dye, and Cohen 2012).

TB is an airborne disease transmitted through infectious contact with an active case. TB transmission is one of the key determinants of epidemic severity, and has important implications for design, implementation, and scaling-up of control interventions (e.g., improved diagnosis, active case finding) that aim to reduce the rate of transmission. Unlike other airborne diseases such as influenza, however, TB transmission is not directly measurable, i.e., available diagnostic techniques cannot estimate the original timing of infection in diagnosed cases. TB also has a predilection for establishment of a latent state that is non-infectious and asymptomatic, but may progress to active, infectious disease at any time. As a result, one cannot reliably differentiate between primary infection with rapid progression to active disease, re-infection following a remote initial infection, or reactivation of a previous latent infection. Such limitations pose several challenges to the study of transmission dynamics across populations, including the lack of informative data to trace the chain of transmission across various contact networks in retrospective studies. The prospective following a cohort population, on the other hand, are prohibitively expensive and restricted by the time and budget constraints. In such settings, the relationships between the duration of disease, symptom burden, contact networks, diagnosis/treatment, and patients’ infectiousness remain obscure (Dowdy, Dye, and Cohen 2012).
We propose an agent-based simulation (ABS) to study TB transmission dynamics and the role of various contact networks. Our model simulates the course of a TB epidemic across a single population and uses a hierarchical network of contacts in three levels, typical to the transmission of airborne diseases (Mossong et al. 2005). Parameters are chosen from the literature, and the model is calibrated to a setting of high TB incidence. We use our model to study the transmission dynamics at an individual level with regard to the timing and distribution of secondary infections from a single source. The average time for disease diffusion to reach 50% of infections at an individual level is estimated, and the timing patterns are compared among different networks. We perform sensitivity analysis of results with regard to multiple parameter values, and discuss the implications for TB control policy.

2 Background

Modeling TB dynamics has a long history, including mathematical models and analytical techniques to describe and predict disease prevalence at the population level (Waaler, Gese, and Anderson 1962; Porco and Blower 1998; Castillo-Chavez and Song 2004). Analytical studies, however, are usually restricted by their simplifying assumptions regarding the population heterogeneity, network structure, and parameters uncertainty, and do not provide a realistic representation of transmission dynamics.

Simulation modeling of TB epidemics in human populations, on the other hand, has a shorter history. One group of studies uses system dynamics to model disease prevalence at the population level (Brewer et al. 1996, Atun et al. 2007). Following a top-down approach, these studies divide the population into different health states, and use transition rates to describe the disease’s natural history. A system of differential equations is used to model disease prevalence through time. In comparison to the analytical approach, such studies
apply a semi-Markov system in which transition rates can change with time, and are able to capture output uncertainty.

Other researchers have developed *discrete-event simulation* (DES) models of TB to evaluate the impact of new diagnostic tools (Langley et al. 2012), or to study more complex structure as in the coinfection of HIV/TB (Hughes, Currie, and Corbett 2006; Mellor, Currie, and Corbett 2011). The DES studies use a schedule of events that are executed in chronological order, and model disease transmission using random generation of Poisson distributions in each mixing group. The aggregate (top-bottom) modeling nature of DES, however, offers low flexibility for direct modeling of contacts (and transmission) at the individual level, and restricts application of such models to certain contact networks.

ABS is an alternative approach that models the system at the individual level and offers high flexibility to incorporate various modeling assumptions. ABS models have wide application in the social sciences (Epstein 1999, Macal and North 2010), and the models have been applied to study various infectious diseases such as influenza (Chao et al. 2010), smallpox (Longini et al. 2006), and HIV/AIDS (Alam, Meyer, and Norling 2008). The literature on application of ABS models in the study of TB, however, is quite limited. Espindola et al. (2011) use an ABS approach to study the emergence of drug-resistant TB due to treatment with antibiotics. They assume a lattice structure for spatial presentation of their population and model TB transmission through local and global interactions across the lattice. The lattice structure, however, does not provide a realistic representation of social settings and cannot be used to study transmission dynamics at the individual level.

We consider a model of transmission involving three contact networks that represent the main social relationships in transmission of an airborne disease. Our model simulates the stochastic contact events at an individual level, and enables us to study patterns of transmission across different networks.
3 Method

Agents represent people in the population whose personal characteristics, social relationships, and health states are programmed at the individual level (Figure 1). The model initiates from an existing epidemic and is calibrated to a setting of high TB incidence. The model is developed using Anylogic (XJ Technologies 2012) and the extensions are coded in Java. A time step of one month is chosen to update the contact networks and model the transmission events, while the internal simulation time-step is very small (0.001 month) and the population dynamics are simulated on a closely continuous time-scale. Outputs are gathered at the end of each year, and the time horizon is 50 years. In this section, we briefly describe the model’s population structure, contact network, and natural history of TB, and discuss our calibration procedure.

![Figure 1: ABS Model Outline](image)
3.1 Population Structure

Following the literature on modeling of airborne infectious diseases (Raffalli, Sepkowitz, and Armstrong 1996) and using a similar approach to Chao et al. (2010), we consider a hierarchically structured population including households, neighborhoods (clusters of households) and communities (sets of neighborhoods). While the exact definition and size of these mixing groups are chosen arbitrarily, their inclusion in the model provides a realistic representation of the population structure, and enables us to model different transmission routes associated with each mixing group.

A household is the smallest and most intimate mixing group for transmission of airborne diseases (Verver et al. 2004, Chao et al. 2010). We assume that household size follows a triangular distribution, ranging from 1 to 10 with a mode of 5 people per family, which corresponds to low socioeconomic status typical of the high-incidence TB settings in India and Brazil. The natural mortality rate is defined at the individual level, and corresponds to a life expectancy of 73 years (Google 2010). We define the probability of child bearing for each family as a decreasing function of the family size such that the average population size remains constant, and the initial household-size distribution keeps its shape through the simulation (see Appendix A).

3.2 Contact Network

We define a three-layer contact network in association with our hierarchical population structure, referred to as close, casual, and random contacts. This is only an abstract representation of real social settings, but the lack of data on TB transmission restricts our ability to use a more sophisticated network. Nevertheless, the three-layer mixing structure enables us to represent the important routes of transmission detected in past TB breakouts.
(Raffalli, Sepkowitz, and Armstrong 1996; Classen et al. 1999), and contrast the dynamics of transmission over each network.

The three contact types represent the social relationships of any individual with the rest of the population (Figure 1). Each type is defined over a specific domain and among a sub-group of the population (e.g. household members). The individual’s probability of engaging in a contact and the contact domain’s size are calibrated to the average number of contacts in each network (Mossong et al. 2005). The contact duration (time delay between network updates representing the persistency of the social bounds) is chosen with regard to the definition of each category (e.g. close contacts are the most persistent type). Contact types are further differentiated by the effectiveness of contacts in transmission of the TB pathogen. The effectiveness parameters are calibrated to provide the given ratios of infections through each network (see Section 3.4).

Close contacts, at the first level, represent the most frequent contact type among household members. On the second level, casual contacts model social relationships such as those among friends at a bar, neighbors at a store, or children at school. These contacts are less frequent and intimate than close contacts, and restricted to a specific network of related individuals including friends, coworkers, etc. (Raffalli, Sepkowitz, and Armstrong 1996; Classen et al. 1999). We restrict the domain of casual contacts to each neighborhood’s residents, and assume a limited period of one year for the duration of contacts (Appendix B). Finally, random contacts are used to model encounters of people at places such as bus stops, museums, etc. Such contacts have the shortest duration (one month), and account for potential risk of transmission among non-related people in the whole community.

The suggested network structure can be further extended to a more detailed network (involving additional mixing groups such as schools, work places, etc.) so that it provides a more realistic view of social settings. The trade-off for developing such models, however, is
lack of suitable data for calibration, as well as the computational expense of more extensive simulation runs.

3.3 TB Natural History

We model TB natural history at the individual level using the five main TB health states (Figure 1). A person is assumed born in full health and susceptible to TB. Upon a successful transmission of the disease, the person enters the *early latent TB* (ELTB) state for a period of five years, during which he endures a high chance of active TB development (*fast progression rate*). At the end of this period, the individual enters the *late latent TB* (LLTB) state, which can last for many years, and is associated with a much lower chance of symptom development (*slow progression rate*). Symptomatic patients with *active TB* (ATB) are infectious and subject to an increased mortality rate. The ATB patients’ infectiousness is modeled as a linear function of time, increasing during the first nine months of the disease from zero to a maximum infectiousness of $I_{max}$, and staying unchanged for the rest of the disease. The $I_{max}$ value is subsequently calibrated to the overall incidence rate (Section 3.4).

Active cases become recovered through spontaneous recovery or upon diagnosis and treatment. Recovered and latently infected individuals are subject to multiple re-infections, upon which they return to the ELTB state. Table 1 shows the disease parameters and their values.
### Table 1: Parameters and Values

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Source/ Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population structure:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of neighborhoods (Number of families per neighborhood)</td>
<td>50 (40)</td>
<td>Calibrated to provide the mean number of contacts in each mixing group (Appendix B)</td>
</tr>
<tr>
<td>Initial household-size distribution</td>
<td>Triangular (1,5,10)</td>
<td>(United Nations Data Retrieval System 2013)</td>
</tr>
<tr>
<td>Life expectancy</td>
<td>73 years</td>
<td>(Google 2010)</td>
</tr>
<tr>
<td><strong>Disease Natural History:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATB mortality rate</td>
<td>0.12 per year</td>
<td>(Tiemersma et al. 2011)</td>
</tr>
<tr>
<td>Early latency duration</td>
<td>5 years</td>
<td>(Dowdy et al. 2012)</td>
</tr>
<tr>
<td>Fast progression rate</td>
<td>0.03 per year</td>
<td>(Vynnycky and Fine 1997)</td>
</tr>
<tr>
<td>Slow progression rate</td>
<td>0.005 per year</td>
<td>(Horsburgh 2004)</td>
</tr>
<tr>
<td>Recovery rate</td>
<td>0.12 per year</td>
<td>(Tiemersma et al. 2011)</td>
</tr>
<tr>
<td>Diagnosis &amp; treatment rate</td>
<td>0.74 per year</td>
<td>Calibrated to provide disease duration of 11 months (WHO 2012)</td>
</tr>
<tr>
<td>Latent immunity toward re-infection</td>
<td>0.8</td>
<td>(Andrews et al. 2012)</td>
</tr>
</tbody>
</table>

### 3.4 Calibration

We model a community of 2000 households in 50 neighborhoods, corresponding to an expected initial population of 10700 people. The community size and its composition is tuned to provide the average number of contacts in each mixing group as described in Appendix B. Individuals are subject to a mortality rate so that some families may eventually die out from the model. We tune the probability of child bearing for each family such that the average population size remains about 10200 people, and the initial family-size distribution keeps a stable shape through time (see Appendix A). The model is initiated from an existing epidemic, and is calibrated to provide an incidence rate of 120 per 100000
people/year, representative of the current global TB incidence rate (WHO 2012). The initial number of infected people at time zero is set to the equilibrium state of a corresponding deterministic model with the same incidence level.

Figure 2 shows the main ABS outputs over 150 years. The *external infection ratio* represents the proportion of disease incidence due to reactivation or progression among individuals who were latently infected at time zero (as a part of the existing epidemic). Since such infections did not originate within the simulation period, they cannot be used for the analysis of transmission. Moreover, due to the stochastic nature of the simulation model and disease dynamics, the annual incidence rate shows a non-stationary pattern during the first decades. We choose the simulation transient period such that the external infection ratio is less than 5%, and the incidence rate has a stationary mean. We test the external infection ratio mean using a one-tailed t-test, and test the stationarity using multiple comparison tests of the incidence mean in consecutive 5-year periods after the transient time (using Bonferroni-adjusted paired t-tests). Consequently, we choose a transient period of 100 years for the simulation model, and start our analysis of the output in year 101.

Figure 2: ABS Output Patterns

Furthermore, we follow the literature on the proportion of TB transmissions in different social settings, and let close contacts account for 20% of transmissions (Verver et al. 2004)
while casual contacts account for 60% of transmission corresponding to 75% of non-household infections (Classen et al. 1999). The effectiveness parameters for each contact type are subsequently calibrated to provide the specified transmission ratios (Table 1). The main outputs of the simulation model including the disease progression measures are provided in Appendix C.

4 Results

We use the calibrated ABS model to study the dynamics of transmission at the individual level and estimate the timing of secondary infections from a single source during the course of the disease. Figure 3a shows the average frequency of infections from a single individual with active TB throughout the course of infectiousness (i.e., from onset of infectiousness, defined as time zero, to treatment or death, the timing of which varies across individuals). The three curves are stratified by the type of contact, and show a similar general pattern: increasing during the initial months as the infectiousness increases, and then decreasing as the infectiousness period ends (e.g., due to recovery). In these figures, the maximum frequency of infections reflects the proportion of transmission due to each type of contact (e.g., higher peak for casual contacts reflects our model assumption that these contacts are responsible for more transmission events than close or random contacts). The different shapes of these curves, however, demonstrate that transmission due to close contact occurs earlier in the course of infection and declines more rapidly over time such that, among individuals with very long periods of infectiousness, random contacts actually become more important sources of transmission than close (household) contacts.

Figure 13b shows the cumulative ratio of infections in each network: half of all transmission events due to close contacts occur within the first 9 months of infectiousness, whereas only about 35% of transmission to non-household contacts has occurred by that
time. Such differences can be partially explained with regard to the relative size of each network and development of immunological protection from repeat infections after initial infection, which limits the number of potential infections. The timing of transmission to casual versus random contacts is remarkably similar, suggesting that, once contact networks reach a certain size (i.e., low probability of reinfection), the timing of infection approaches a limiting shape regardless of the assumed duration and intensity of the contact network.

Figure 3: Timing of Secondary Infections from a Single Source

Figure 4 shows the distribution of secondary infections in each network, stratified by type of transmission to an uninfected individual (new infection), versus a previously infected individual (re-infection). New infections show the earliest increase among close contacts due to the high frequency of household contacts. However, as more household members become infected, the number of new infections rapidly declines, and the number of re-infections increases, such that more household infections represent re-infections than new infections after about 12 months of infectiousness, and this ratio continues to grow over time (Figure 5). In contrast, casual and random contacts rapidly reach a plateau ratio of new infections to re-infections that remains relatively low, reflecting a large pool of susceptible individuals relative to the number of transmissions occurring.
Finally, we performed a series of one-way sensitivity analyses to ascertain the effect of varying each individual parameter on the time to reach 50% of transmissions. We varied the duration of casual contacts from 1 to 60 months, the relative sizes of non-household networks by a factor of two in either direction, and the ratio of transmissions in each network (using a three-level factorial experiment with two factors to cover various settings). In each scenario, the effectiveness of contacts were recalibrated to provide the specified ratios of transmission, and the $I_{max}$ value is tuned to provide the incidence level of 120 per 100000/year. The results show that the time to reach 50% of transmissions is robust to wide but not-unreasonable variation of most parameters, and the only notable effect was observed from increasing the household transmission ratio to a very high level (80%), which resulted in a longer duration of 11 months to reach 50% of transmission in this network.
5 Summary and Discussion

We propose an ABS approach to study the timing and distribution of TB transmissions through the course of the disease. We consider a hierarchical contact network at three levels (close, casual, and random) to represent the important social settings identified in the transmission of airborne infectious disease and past TB outbreaks. The model is calibrated to a setting of high TB incidence, and the proportion of transmission through each network was adopted from the literature.

A main implication of the timing and distribution of transmissions is for the development of TB diagnostic interventions. Our results suggest that nearly half of all TB transmissions occur after an individual has already been infectious for over one year. Therefore, implementing systems that facilitate diagnosis and treatment of cases in the first year of disease (e.g., annual screening of target populations) can avert a substantial fraction of a population’s TB transmission burden.

Moreover, our results show that household (close) transmission events occur more quickly in time than transmissions to casual/random contacts. Thus, household-based interventions (e.g., contact tracing) need to occur relatively soon in the infectious course to have maximum impact, whereas population-based interventions (e.g., population screening) may still have important impact if they are widely spaced.

Finally, we detected a remarkably similar pattern for timing of transmissions between non-household contacts, suggesting that there may be a common “shape of transmission over time” outside the household; transmissions to non-household contacts may occur with roughly similar timing regardless of the intensity and duration of contact. This finding may help to simplify future efforts at TB simulation and data collection, by relaxing the
requirement to define wide arrays of contact networks, focusing instead on two strata of household vs. non-household (or close vs. other) contacts.

Our work is limited by a relative lack of empirical data on TB transmission and parameter values in the associated contact networks. Our sensitivity analysis of outputs indicates a robust behavior of the results to variation of main parameter values. Nevertheless, additional studies are required to assess the impact of various network structures on timing of transmissions.

Future work include extension of the model to more realistic settings involving multiple interacting communities, age-structured populations, key risk factors (e.g., malnutrition, smoking, poverty), coinfections (especially HIV), and interventions (e.g., household contact tracing) that depend on contact networks and agent’s behavior. Subsequent models can also address various policy-making issues such as optimal allocation of TB control resource among multiple populations and interventions (Kasaie and Kelton 2012a, 2012b).

6 References


Supplementary Material

1. Family Probabilities of child bearing

A family’s probability of child bearing at each month is defined as a decreasing function of the family size, so that the average population size remains constant, and the initial household size distribution keeps a stable shape through the course of the simulation. Using an empirical exponential function (dotted line, Figure 6) to describe the inverse relationship, we tune the individual probabilities for each family size by trial and error. Figure 6 (solid line) shows our final empirical function modeling the decreasing relationship of family size and the probability of child bearing.

![Figure 6: Probability of Child Bearing for Families](image)

2. Contact Network calibration

Transmission of TB occurs due to contact with an infectious individual, but the probability of transmission depends on several factors, including the duration of exposure, infectiousness of the case, immunity of the contacted individual, and environmental factors
such as the ventilation system. In such a setting, close contacts are assumed to have the highest probability of transmission, while casual contacts have a lower probability. Households are usually regarded as places of close contact and high risk for transmission of TB, while the remaining contacts outside of the household are referred to as casual. However, the exact definition of a close contact is not clear, and the relative importance of crowding or contact with a TB patient outside the household has not been established.

Classen et al. (1999) conduct a prospective study in a high-incidence TB setting, and assess the role of social interactions in transmission of TB. Their results show that in such high-incidence areas, contacts outside of the household play an important role in transmission, and accounts for 70% of transmissions. Such contacts usually take place during recreation, which in this setting consist of drinking in social groups. These events usually take place at a local neighborhood informal bar or at people’s homes, with an average frequency of 4-5 times per week.

In this sense, we define our contact network in three levels representing the close contacts inside households, and the casual and random contacts outside households. In this definition, random contacts represent short-term encounters with non-related individuals in settings such as transportation, exhibitions, etc. Casual contacts, however, represent a more stable relationship with a smaller population of friends, co-workers, neighbors, etc. who are contacted multiple times for the duration of the relationship. We allow each individual to form his or her networks of contact with a certain likelihood at each time step (e.g., a person may choose to have no random-contact at a given month), and once a network is formed, the members continue to contact each other.

In order to determine the size of each network, we refer to Mossong et al. (2008), who conducts a prospective study on mixing patterns and contact characteristics relevant to the spread of infectious disease. Their results suggest a mean value of 13.4 daily contacts for
each person, with 20% of contacts’ occurring at home. We therefore estimate a frequency of 11 daily contacts outside of households, corresponding to 330 contacts in each month. Consequently, we calibrate out contact network parameters as follows:

We let the likelihood of attendance represent the individual’s likelihood of forming a contact network. Once a network is formed, it lasts for the duration of contact and the mutual contacts among the network members are executed at each time step. We assume that the probability of attendance in close contacts is 1 (all family members are engaged in a contact), and estimate the likelihood of attendance in casual contacts from Classen et al. (1999). Moreover, we assume that the close contacts have infinite duration (so that the family members are always in contact with each other), while the duration of casual contacts is one year (12 steps) and random contacts last for one month. With regard to the definition of each network, we restrict the domain of close contacts to household members, and casual contacts to neighborhood residents. We let the total casual and random network size be around 330 people (Mossong et al. 2005), and assign about 35% of these contacts to the casual network (Classen et al. 1999). These assumptions are then used to calibrate the number and size of neighborhoods in our model. Table 2 shows a list of network parameters and their values in the simulation model.

Table 2: List of Contact Network Parameters

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Contact Type</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Close</td>
<td>Casual</td>
<td>Random</td>
</tr>
<tr>
<td>Domain</td>
<td>Household</td>
<td>Neighborhood</td>
<td>Community</td>
</tr>
<tr>
<td>Duration</td>
<td>Forever</td>
<td>1 y</td>
<td>1 Month</td>
</tr>
<tr>
<td>Target Network Size</td>
<td>Household Size</td>
<td>210</td>
<td>90</td>
</tr>
<tr>
<td>Likelihood of Attendance</td>
<td>1</td>
<td>0.65</td>
<td>0.15</td>
</tr>
<tr>
<td>Domain Size (Mean)</td>
<td>Household Size</td>
<td>200</td>
<td>Entire population</td>
</tr>
<tr>
<td>Estimated Network size in model</td>
<td>Household Size</td>
<td>84.5</td>
<td>229.5</td>
</tr>
<tr>
<td>Effectiveness of contacts</td>
<td>1</td>
<td>0.09</td>
<td>0.009</td>
</tr>
</tbody>
</table>
3. Simulation Output in the Base scenario

Table 3 shows the main outputs of the simulation model.

Table 1: ABS Outputs

<table>
<thead>
<tr>
<th>Output</th>
<th>Mean +/- 95% Half-Width Interval</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidence</td>
<td>12.24 +/- 0.33</td>
<td>people/year</td>
</tr>
<tr>
<td>Prevalence</td>
<td>11.2 +/- 0.33</td>
<td>people/year</td>
</tr>
<tr>
<td>Duration</td>
<td>10.94 +/- 0.05</td>
<td>year</td>
</tr>
<tr>
<td>Fast Progression</td>
<td>10.33 +/- 0.31</td>
<td>people/year</td>
</tr>
<tr>
<td>Slow Progression</td>
<td>1.9 +/- 0.04</td>
<td>people/year</td>
</tr>
<tr>
<td>ATB Mortality</td>
<td>1.36 +/- 0.04</td>
<td>people/year</td>
</tr>
<tr>
<td>Recovery</td>
<td>1.34 +/- 0.04</td>
<td>people/year</td>
</tr>
<tr>
<td>Treatment</td>
<td>9.37 +/- 0.26</td>
<td>people/year</td>
</tr>
</tbody>
</table>
Essay 2: Timing of Tuberculosis Transmission and the Impact of Household Contact Tracing: An Agent-Based Simulation Model

Parastu Kasaie¹, Jason R. Andrews², W. David Kelton¹, David. W. Dowdy³

¹ Department of Operations, Business Analytics, and Information Systems, College of Business, University of Cincinnati, Cincinnati, Ohio 45221, USA.
² Division of Infectious Diseases, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
³ Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21218, USA.

Abstract

Household contact tracing has recently been endorsed for global tuberculosis (TB) control, but its potential population-level impact remains uncertain. Our objective is to project the maximum impact of household contact tracing for TB in a moderate-burden setting. We developed a stochastic, agent-based simulation model of a simplified TB epidemic, calibrated to a setting of moderate TB incidence. We used data from the literature to generate “community-driven” and “household-driven” scenarios in which 22% and 50% of TB transmission occurred within the household respectively. In each scenario, we simulated an intervention in which the household members are screened and treated for TB at the time of an index patient’s active TB diagnosis. Our results suggest that by the time of TB diagnosis, 75-95% of initial household infections had already occurred, but only 1.5-3.0% of contacts had sufficient time to progress to active TB. With 100% sensitive tracing of all contacts for five consecutive years, TB incidence declined by 10-15%; a mean year-over-year decline of 2%/year. Effects were sustained for many years after stopping the intervention. Providing preventive therapy with contact tracing nearly doubled this impact (17-27% decline in incidence). Impact was proportional to sensitivity and coverage; thus, if 50% of contacts were screened with a 50% sensitive test, TB incidence declined by only 0.5%/year. Household contact tracing is unlikely to transform TB epidemiology in isolation but has the potential – especially with provision of preventive therapy – to augment a comprehensive package of interventions that could substantially reduce the population-level burden of TB.

Key Words: tuberculosis; contact tracing; models, theoretical; epidemiology
1 Introduction

As long-term goals for tuberculosis (TB) control are being developed, focus is increasingly shifting to ambitious targets, such as elimination of TB by 2050 (Lonnroth et al. 2013; Dye et al. 2013). Although TB incidence is falling by 2-3% per year worldwide (World Health Organization 2013b), traditional control strategies are unlikely to hasten this rate of decline dramatically (Dowdy & Chaisson 2009). As a result, screening for active TB in high-risk groups is emerging as one strategy for enhanced population-level TB control (Lonnroth et al. 2013; Kranzer et al. 2013). Among screening strategies, household contact tracing (HHCT) (Fox et al. 2013) – screening and treating the household members of people diagnosed with active TB – is attractive because it is logistically feasible and provides a reasonably high yield of TB cases. However, while a number of analyses have evaluated the population-level impact of interventions including community-based active case finding (Corbett et al. 2010; Ayles et al. 2008), improved diagnostic testing (Cohen et al. 2012; Dowdy et al. 2006), and novel drugs or vaccines (Abu-Raddad et al. 2009; Murray & Salomon 1998), the potential population-level impact of household contact tracing is less clear (Begun et al. 2013). This partially reflects our lack of understanding of when TB transmission occurs within households and similar network structures (Mills et al. 2011; Cohen et al. 2007). While estimates exist of the amount of TB transmission that occurs within households (Verver et al. 2004; Madico et al. 1995), the timing of that transmission in relation to contact tracing is less clear. On the one hand, it is possible that the majority of household transmission occurs well before contact tracing can take place (Dowdy et al. 2013); thus contact tracing may be too late in the transmission/disease cycle to reduce incidence unless secondary (and tertiary) cases can also be identified. On the other hand, it is also possible that contact tracing occur before household members have sufficient time to
develop active TB. These alternatives have important implications for the evaluation and design of existing contact-tracing programs.

The population-level impact of household contact tracing is difficult to study empirically because of the large sample size required for adequate power. For example, the ZAMSTAR study (Ayles et al. 2008) covered a population of nearly one million individuals and found an epidemiologically important 18% decline in TB incidence over four years comparing household intervention to a control, but this difference was not statistically significant (Ayles et al. 2008). Disease-transmission models can be used to understand and project the likely impact of policy options for disease control when empirical data are difficult to collect at the population level (Epstein 2009; Kasaie & Kelton 2013b). In particular, agent-based simulation enables inference at the individual level (e.g., transmission between cases of active TB and their household contacts), and enables realistic implementation of disease-control interventions such as contact tracing (Epstein 1999; Guzzetta et al. 2011; Begun et al. 2013). Thus, to estimate the timing of household transmission and evaluate the likely impact of household contact tracing for TB, we constructed an agent-based simulation model of TB transmission in a simplified population with explicit household structure (see Section 1 in the Supplementary Material). Formative research on which the current study is based has been previously reported in (Kasaie et al. 2013).

2 Methods

2.1 Model and Rationale

We developed an agent-based simulation model of a TB epidemic in a simplified population of 2,000 households (Figure 1). The household sizes vary from 1 to 10, corresponding to a discrete triangular distribution with a mode of 5 people (representing a typical household
structure in India and Brazil (System 2013)). Given the paucity of existing models of household contact tracing in TB (Begun et al. 2013), we sought to construct the simplest possible formulation that would yield a generalizable inference; as such, we do not explicitly incorporate such elements as age structure (as pediatric TB is not infectious) or HIV (as empirical data on the natural history of untreated HIV-associated TB are sparse and conflicting). We reason that a simple model is required to understand “first principles,” and once those principles are elucidated, future models can incorporate additional structure to provide further insight.

![Diagram of TB model]

**Figure 1: Agent-Based Model Outline.**

In this model, people are modeled according to their TB status (left panel) and membership in a household of random size from 1 to 10 (right panel). Individuals with active TB interact much more closely and frequently with other members of their household (household contacts) than with members of other households (casual contacts). The individuals are subject to the natural mortality rate (not shown) at all times. *The ELTB state is composed...
of 5 one-year sub-states, each associated with a specific rate of fast progression as noted in Table 1

2.2 Network and Population Structure

Individuals in this model are described in terms of their household membership and TB status. We model two sources of disease transmission, among the household members (close contacts) and the community members (casual contacts). Each contact type is associated with a frequency and effectiveness parameter that determines the likelihood of disease transmission in each network (see Section 1.3 in the Supplementary Material). Individuals are assumed to contact all of their household members and selected members of their community at a time step of one month. The selection of a one-month time step is long enough to facilitate crude estimation of monthly social-interaction patterns (given the scarcity of data on daily contact patterns associated with airborne – rather than droplet-borne – transmission), yet short enough to reveal the dynamics of TB transmission across a year-long period. Regarding social interaction, we estimated the average number of casual contacts (outside the household) in each month using the results of a relevant study for spread of respiratory infectious disease (Mossong et al. 2008) (Table 1), while noting as a limitation that TB – as an airborne disease – may have very different transmission characteristics from those (e.g., influenza) spread by droplets. The contact rate between an active case and his/her household members (close contacts) is larger than that to other community members (casual contacts), and is calibrated to provide the specified household transmission ratio in each scenario. Greater detail about the simulation process is provided in Section 1 in the Supplementary Material.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Basecase Value</th>
<th>Sensitivity Analysis {Low, High} Limits</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population structure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household size mean (range)</td>
<td>5 (1-10)</td>
<td>{2.5 (1-5), 10 (1-22)}</td>
<td>(System 2013)</td>
</tr>
<tr>
<td>Life expectancy</td>
<td>73 years</td>
<td>{40, 80} years</td>
<td>(Google 2010)</td>
</tr>
<tr>
<td>TB Natural History</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATB mortality rate</td>
<td>0.12 /year</td>
<td>{0.05, 0.4} /year</td>
<td>(Tiemersma et al. 2011)</td>
</tr>
<tr>
<td>Early latency duration</td>
<td>5 years</td>
<td>{2, 5} years</td>
<td>(Holm 1969)</td>
</tr>
<tr>
<td>Cumulative fast progression rate</td>
<td>14.30%</td>
<td>{5-17} % /5years</td>
<td>(Vynnycky &amp; Fine 1997)</td>
</tr>
<tr>
<td>Annual progression risk during</td>
<td>(8.66, 3.55, 1.12, 0.74, 0.24)%</td>
<td></td>
<td>(Vynnycky &amp; Fine 1997)</td>
</tr>
<tr>
<td>Early latency (year 1 to 5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slow progression rate</td>
<td>0.0005 /year</td>
<td>{0.0002, 0.001} /year</td>
<td>(Horsburgh Jr 2004)</td>
</tr>
<tr>
<td>Recovery rate</td>
<td>0.12 per year</td>
<td>{0.05, 0.4} /year</td>
<td>(Tiemersma et al. 2011)</td>
</tr>
<tr>
<td>Treatment rate</td>
<td>0.74 /year</td>
<td>{0.5, 2.0} /year</td>
<td>(World Health Organization 2013b)</td>
</tr>
<tr>
<td>Latent immunity toward re-infection</td>
<td>80%</td>
<td>{20%, 80%}</td>
<td>(Andrews et al. 2012)</td>
</tr>
<tr>
<td>Incidence rate</td>
<td>120 per</td>
<td>{20, 300} per</td>
<td>(World Health Organization 2013b)</td>
</tr>
<tr>
<td>Disease duration until maximum infectiousness attained</td>
<td>9 months</td>
<td>{6 ,12} months</td>
<td></td>
</tr>
</tbody>
</table>
2.3 Natural History of TB

TB natural history is modeled at an individual level using five main TB states (Figure 1), following other models of TB (White & Garnett 2010): uninfected, recently infected within the last five years (“early latent” TB, ELTB), remotely infected (“late latent” TB, LLTB), active TB disease (ATB), and recovered (REC). Individuals with latent TB are assumed to be non-infectious, and those with ELTB have an increased probability of progression to active disease relative to those with LLTB. This probability declines with each year in the ELTB state (i.e., the ELTB state consists of five sub-states, each lasting for one year), according to estimates of adult TB progression following infection (Vynnycky & Fine 1997). Reinfection is modeled as a return to the first year of the ELTB state.

Patients with active TB are infectious and subject to an increased probability of mortality. Infectiousness is modeled as zero at the start of the active period, increasing thereafter as a linear function of time for the first nine months of disease (i.e., as bacillary burden grows), and stable thereafter at the maximum level until the individual is diagnosed and treated, or dies. The maximum transmission probability is calibrated to provide the specified incidence rate at baseline. Individuals recover from active TB by diagnosis and initiation of effective treatment, or alternatively through spontaneous resolution. We calibrate the treatment rate such that individuals remain infectious for an average of 11 months prior to treatment, the global mean as estimated by the World Health Organization (World Health Organization 2013b). These individuals are also subject to a risk of reinfection, as described above.

2.4 Calibration

Model parameters were obtained from the literature (Table 1), with the per-contact transmission risk calibrated to provide an incidence of 120 per 100,000 population/year (the
global average incidence (World Health Organization 2013b)). In each experiment, sufficient model runs were performed (more than 1000) to achieve a relative precision of at worst 1% for the incidence rate (i.e., the incidence rate falls between 120 ± 1.2 per 100,000/year). Since the proportion of TB transmission that occurs in households (vs. the general community) remains a matter of uncertainty, we constructed two alternative and data-driven scenarios. In the first (community-driven) scenario, which may be more reflective of settings with very high incidence, we follow results from South Africa by Verver et al. (2004) and assume that the proportion of TB infections due to household contact is 19% (Verver et al. 2004). In the second (household-driven) scenario, which may be more reflective of moderate-incidence settings, we use estimates from Peru by Brooks-Pollock et al. (2011) in which each case of active TB is assumed to infect 50% of his/her household members before treatment or death (Brooks-Pollock et al. 2011). In our model, this assumption corresponds to 32% of all initial infections and 50% of all transmission events (initial infections and reinfections) occurring within the household (Supplementary Material, Section 1.3.). Modeling and analysis were performed using AnyLogic (XJ Technologies 2013); source code and a demonstration are publicly available at: http://www.runthemodel.com/models/k-X4D57KfPBfauXuC16q2Z/. A complete list of model outputs under the two scenarios is provided in Section 2 of the Supplementary Material.

2.5 Interventions and Outcomes

After constructing our calibrated populations, we introduced a household contact tracing intervention, in which a proportion of passively diagnosed TB cases (“index cases”) are selected. These individuals’ household members are then screened and treated for active TB; in some cases, we also consider preventive therapy (PT) that is capable of reducing the
future risk of reactivation (though not reinfection) by 70% (Ferebee 1969; Akolo et al. 2010). In order to estimate the maximum impact achievable with contact tracing, we assume tests with 100% sensitivity for diagnosing active TB, but we also assess the relationship between impact and lower levels of population coverage or diagnostic sensitivity. For purposes of comparison, we also considered a community-based active case finding (ACF) strategy, in which a randomly-selected set of households is chosen at annual intervals, all members are screened for TB, and the diagnosed active cases are treated immediately. Our primary outcome of interest was the percent reduction in TB incidence achievable by each intervention at the end of a sustained five-year campaign. For comparability, the ACF strategy is calibrated to provide the same incidence reduction in the community-driven scenario. We performed one-way sensitivity analyses around key model parameters in Table 1, and reported the estimated results from multiple simulations.

3 Results

3.1 Timing of Infections

When we calibrated the treatment rate to provide an average duration of 11 months for untreated active disease, over 50% of infections to casual contacts nevertheless originated from individuals who had been infectious for more than 11 months (Figure 2B, dashed lines). By contrast, even though we assumed lower infectiousness during the first 9 months of active TB, the majority of new infections to household members occurred within 4.0 months (household-driven scenario, light-grey line) to 7.0 months (community-driven scenario, dark-grey line) of infectious onset. By the time of contact tracing, 75% (community-driven) to 95% (household-driven) of new household infections had already occurred. The rate of household infections peaked three to five months sooner than that of casual contacts, with more intense household infection resulting in faster “saturation” of
household contacts (Figure 2A). Under all explored parameter variations, the median household infection consistently occurred 1.5-3 months before the median infection among casual contacts (see Section 2.1 of the Supplementary Material for further discussion).

Panel A shows the number of new infections (i.e., transition from susceptible to early latent TB, similar to a tuberculin skin-test conversion, excluding reinfections) per month occurring from an average case to close contacts (solid) and casual contacts (dashed); panel B shows the proportion of all new infections that occur by a given month in the infectious period, assuming an average duration of infectiousness of 11 months. Light-grey lines show a “household-driven” scenario in which 50% of transmissions occur within the household; dark-grey lines denote a “community-driven” scenario in which 22% of transmissions occur within the household. Early in the disease course, the monthly number of infections rises as infectiousness increases (a model assumption); over time, this number decreases as the

Figure 2: Timing of secondary new infections from a single case, assuming an infectious period of mean 11 months.
probability of spontaneous cure, death, or treatment rise and the number of susceptible contacts in the household declines. Note that the dashed lines overlap in panel B.

### 3.2 Case Finding: Programmatic Indicators

If all household contacts in a community were successfully traced using a diagnostic test with perfect sensitivity, our model projected that a total number of 1950 (± 31) contacts per 100,000 population would be identified in the community-driven scenario (over five years), of whom 620 (32%) would have early latent TB infection (infected in the last 5 years), 610 (31%) would have late latent TB infection (infected more than 5 years in the past), and 30 (1.6%) would have active TB. Corresponding proportions in the community were 3% early latent TB, 36% late latent TB, and 0.1% active TB (i.e., TB prevalence near 100 per 100,000). Thus, tracing 2,000 contacts would identify as many cases of active TB (and nearly as many early latent infections) as community-based screening of 30,000 individuals (Supplementary Material, Section 2.2).

### 3.3 Case Finding: Impact on TB Incidence

Assuming that every household contact with active TB in the community-driven scenario could be identified, diagnosed, and treated, TB incidence would fall 10% by the end of five years (from 120 [± 2] per 100,000/year to 108 [± 1.9] per 100,000/year), as shown in Figure 3 (HHCT dark-grey bar). In this scenario, TB incidence did not fall in year 1, fell by only 1% in year 2, and then fell by a mean of 3%/year in years 3-5: a mean year-over-year decline of 2%/year. After the intervention was stopped at the end of year 5, incidence continued to decline for two more years and stayed below the baseline level for at least 15
years after the intervention stopped (Figure 4). A similar pattern was observed for the household-driven scenario.

Each scenario compares a strategy with 100% diagnostic sensitivity and 100% treatment initiation among those found to have TB (active TB only in the two strategies on the left, latent TB also in the strategy with preventive therapy, PT, on the right) against a baseline scenario in which no case-finding intervention is undertaken. Numbers represent the average incidence level (± half-width 95% confidence interval around mean) in each scenario. The primary outcome is the percent reduction in adult TB incidence at the end of five years. Interventions include: household contact tracing (HHCT) for 100% of treated cases, annual community-based active case finding (ACF) with 35% population coverage, and household contact tracing for 100% of treated cases with preventive therapy of latently-infected
contacts (HHCT+PT). Realistic interventions using diagnostic tests with suboptimal sensitivity and incurring losses to follow-up will perform less well than depicted here.

![Graph showing incidence changes under HHCT program]

**Figure 4: Year-to-year incidence change by household contact tracing strategy.**

This figure shows the year-to-year incidence changes under a HHCT program lasting five years in each scenario. The HHCT program is initiated at the beginning of the first year and continues for 5 consecutive years. The percent reduction in incidence was computed at the end of intervention (in year 6) and was used as the primary outcome in our analysis. In both scenarios, the TB incidence showed a small reduction in the first two years, and then fell by a mean of 3%-4%/year in years 3-5: a mean year-over-year decline of 2%-3%/year. After the intervention was stopped at the end of year 5, incidence continued to decline for two more years and did not return to the pre-intervention baseline for over 20 years. The reported average incidence levels are subject to stochastic variation with an average of 2 people per 100,000/yr.

To achieve the same 10% decline in incidence at the community-driven scenario, 35% of individuals would have to be screened through an active case-finding strategy over the
five-year period (Figure 3, ACF dark-grey bar). In the household-driven scenario (relative to the community-driven scenario), performing contact tracing had 50% more impact on incidence (15% vs. 10% reduction in five-year incidence from complete contact tracing), whereas active case-finding had 30% less impact (7% vs. 10% reduction from screening 35% of the population, Figure 3 ACF bars). For all interventions, transmission impact was a nearly linear function of the diagnostic test sensitivity and the proportion of screened TB cases who were effectively diagnosed and treated (Supplementary Material, Section 2.3). Thus, for example, if only 50% of contacts were screened with a 50% sensitive test, the projected impact on TB incidence was only 50% * 50% = 25% as great (i.e., mean year-over-year decline of only 0.5%/year). Since the number of individuals with early latent TB identified by household contact tracing exceeded the number of people identified with active TB by a factor of 20, the addition of preventive therapy (PT) increased the impact of household contact tracing on TB incidence by 1.7-1.8 fold regardless of the intensity of household transmission (Figure 3, HHCT+PT bars).

### 3.4 Sensitivity Analysis

The impact of household contact tracing on TB incidence was robust to changes in most model parameters (Figure 5), including household size, duration of infectiousness (i.e., speed of passive diagnosis and treatment), and duration of the high-risk “early latent” period (see Section 2.1.4 of the Supplementary Material for more information on the experimental design). Transmission impact of contact tracing was sensitive to the proportion of disease due to recent infection, having least impact in low-incidence settings where the majority of incident TB was due to reactivation. Transmission impact increased almost linearly with increasing incidence until the underlying TB incidence reached a level of 50 per 100,000/year, above which level the vast majority of incident TB was due to recent
transmission; beyond this level, the impact of contact tracing decreased somewhat as the ratio of household transmission decreased through time (assuming that higher-incidence settings have a higher proportion of community-based transmission, see Section 2.4 in the Supplementary Material). Results were also sensitive to the timing pattern of progression to active TB during the early latent period; assuming that progression risk was equal across all five years (e.g., 2.9% per year) rather than concentrated in the first year after infection (as in the base model) decreased the five-year impact of HHCT on incidence from 10% to 3%.

This figure shows the sensitivity analysis of the intervention’s effectiveness to variation of model’s parameter values in the community-driven scenario (22% of TB transmission in the household), with a baseline incidence level of 120 per 100,000 population per year (except for the incidence variation scenarios; see Section 2.1.4 in the online data supplement). We tested all parameters in Table 1 but excluded parameters that caused <5% variation in either
direction, as these small fluctuations may reflect simulation stochasticity rather than true behavior. The data labels show the value of incidence reduction in each case, relative to the household contact tracing at baseline (10%). For three variables, variation both above and below the baseline value resulted in change to the primary outcome (percent incidence reduction relative to baseline) in the same direction. In these cases, we show the variation causing the larger deviation from the baseline estimate (10%). *When incidence was increased to 300 per 100,000/year, the percent reduction in incidence fell to 7.5% (not shown).

4 Discussion

This household-structured, agent-based simulation model elucidates the transmission dynamics of TB within households and the likely epidemiological impact of household contact tracing. Specifically, our model suggests that, whereas the majority of community-based transmission likely occurs from individuals who have been infectious beyond the mean infectious duration of TB (over a year in most high-burden settings), most new infections within the household may occur in less than half that time. In this simulation model, the maximum five-year reduction in TB incidence achievable by household contact tracing was 10-15% (2-3% per year) with proportionally lower impact with imperfect coverage or sensitivity. However, TB incidence continued to decline for two years after program cessation and remained below baseline levels for over 15 years after a five-year contact tracing intervention; addition of preventive therapy nearly doubled this impact. These results provide important insight into the role of household structure in TB transmission and may help program officials more accurately project the medium-term epidemiological impact of contact-tracing interventions.
Although our results were simulated in a simplified population, they conform well to data gathered in field settings. For example, we estimated that 1.5-3.0% of household contacts would have active TB, a proportion that corresponds closely to the proportion of bacteriologically-confirmed TB (2.3%) in a 2008 meta-analysis (Morrison et al. 2008). Our estimates of the proportion of latent infection among household contacts (63%) also correspond well with the 51.4% reported in that analysis, assuming a 77% sensitivity of tuberculin skin testing for latent TB infection (Morrison et al. 2008) (i.e., true proportion of latent TB 51.4%/0.77 = 67%). Thus, the proportional burden of active and latent TB in our simulated households corresponds well to actual population-based data.

Currently, limited evidence exists to demonstrate convincingly that case finding interventions – including household contact tracing – reduce TB incidence at the population level (Kranzer et al. 2013). The ZAMSTAR study (Ayles et al. 2008) – a large community-randomized trial of active case finding in Zambia and South Africa – found an 18% reduction in TB prevalence in its household-based arm but no impact on prevalence from an (arguably more intensive) “enhanced case finding” intervention. Our simulation results demonstrate why this might be the case for household contact tracing. First, a 2-3% reduction in incidence per year, though capable of creating important declines in TB incidence over time, is unlikely to be detected in any real-world study with imperfect coverage and sensitivity (and falls well within the confidence intervals of both arms of the ZAMSTAR study). Second, the impact of contact tracing on transmission is delayed – although our modeled intervention eventually achieved a 3-4% year-over-year reduction in TB incidence, we saw virtually no effect for the first two years following implementation. Studies seeking to evaluate the transmission impact of contact tracing must therefore follow population trends for a number of years to estimate the true population-level impact of such interventions. Finally, providing preventive therapy to latently infected adult contacts (not
routinely done in most evaluations of contact tracing) nearly doubles the overall potential impact of the intervention. Thus, household contact tracing may generate important reductions in TB incidence (up to 7% decline year-over-year if combined with preventive therapy), but this impact is unlikely to be seen in short-term evaluations of incomplete contact tracing interventions without universal preventive therapy. Given that household contact tracing is far more efficient than population-based active case finding – by a factor of 15 or more (in terms of the number needed to screen to find one case of active TB) – greater effort is warranted to evaluate the long-term impact of contact tracing and to optimize its population-level benefit through corresponding provision of preventive therapy.

As with any model-based analysis, this agent-based simulation has a number of limitations. First, given the dearth of modeling investigations in this area and the paucity of data to parameterize a more complex model, we assumed a population with a “global average” TB incidence and without age structure or HIV. As a result, quantitative results from this model are unlikely to generalize precisely to specific settings, particularly those with high HIV burden. Children and HIV-infected individuals are critical targets of household contact investigations due to their high mortality risk after TB infection; thus, our model likely underestimates the impact of household contact tracing on TB mortality. However, as these two groups contribute less to TB transmission on a per-person-year basis (i.e., are more likely to have smear-negative or extrapulmonary TB), our estimates of transmission impact (the primary focus of this analysis) may be accurate. Second, our model does not explicitly incorporate migration and therefore cannot be directly generalized to low-incidence settings (e.g., Western Europe, United States, Canada) with high proportions of TB disease found in foreign-born populations. Third, we adopted a number of simplifying assumptions, including lack of spatial correlation between households and a linear increase in per-contact transmission rate from onset of infectiousness to plateau levels; more
complex modeling efforts could further explore the role of these assumptions in modifying results, which in turn highlights the need for better data on TB transmission. Moreover, we assumed perfectly sensitive screening for active and latent TB, but diagnostics may be less sensitive particularly early on in the course of active TB. Similarly, tests for latent tuberculosis may be negative in the weeks to months following initial infection, precluding preventive therapy if a positive test for latent TB is required. Our results therefore demonstrate an upper bound for the impact of the case-finding strategies. Finally, we did not explicitly incorporate costs or resource requirements of contact tracing; the comparative economics of TB control is also an important area of future study.

In summary, we have used an agent-based simulation model to explore the dynamics of TB transmission within vs. across households, and to inform estimates of the population-level impact of household contact tracing. We find that contact tracing can have substantial epidemiological impact (up to 7% reduction in incidence per year), but only if it achieves relatively complete population coverage, is sustained over time, and includes preventive therapy of latently infected contacts. Since the transmission impact of contact tracing is lagged by approximately three years, short-term evaluations of contact tracing are likely to underestimate their long-term impact substantially. Ultimately, household contact tracing is unlikely to have transformative impact on TB epidemics in isolation but can generate epidemiologically important reductions in transmission, especially if combined with preventive therapy. Contact tracing should be subjected to longer-term evaluation while being included as an important component of a comprehensive package of TB control interventions that, taken together, can hasten progress toward global TB elimination.
5 References


Holm, J., 1969. Development from tuberculous infection to tuberculous disease. TSRU Progress Report;KNVC; The Hague, The Netherlands,


Supplementary Material

1 Household-Structured Simulation Model of TB

While compartmental ordinary differential equation models have been commonly used to describe transmission dynamics of tuberculosis (Porco & Blower 1998; White & Garnett 2010; Castillo-Chavez & Song 2004), other individual-based simulation techniques including agent-based simulation (ABS) have been less widely applied (de Espíndola et al. 2011; Mellor et al. 2011; Guzzetta et al. 2011). Such models allow for a more realistic implementation of interventions (e.g., household contact tracing) whose impact depends on individual contact networks (Begun et al. 2013). Whereas ABS has been widely applied to other infectious diseases (Epstein et al. 2007; Longini et al. 2007; Longini et al. 2005; Burke et al. 2006; Epstein 2009; Chao et al. 2010), its applications in TB have been limited to only a small number of studies (de Espíndola et al. 2011; Guzzetta et al. 2011; Cohen et al. 2007), due primarily to the scarcity of available data on contact networks in TB, as is required to calibrate and validate more complex models (Chao 2013; Dowdy, D. W., C. Dye et al. 2013).

In this study, we develop an ABS model of TB transmission in a simplified population with explicit household structure. Unlike other infectious diseases (e.g., influenza (Eames et al. 2012; Glass & Glass 2008) or sexually transmitted diseases (Wiley et al. 1989; Jacquez et al. 1988; Latora et al. 2006)), data on the intensity and frequency of “contact” in TB are scant due to its airborne route of transmission, variable incubation/latency period (from a few months to many years), and lack of reliable markers (e.g., antibody titers) of recent exposure and transmission. Thus, unlike models of other diseases for which complex transmission networks can be parameterized (Cauchemez et al.}
models of TB require simplicity, with a minimum of assumptions and data requirements, for meaningful inference.

The disease parameters are calibrated to the epidemiological literature, reflective of a moderate incidence setting. In order to study the timing of TB transmission in households and the effectiveness of household contact tracing, we consider a heterogeneous population with a two-level mixing structure (Kasaie et al. 2013). Agents represent people in the population whose personal characteristics, social relationships, and health states are programmed at the individual level. The model is developed using the Anylogic simulation software (XJ Technologies 2013), and the extensions are coded in Java.

A timestep of one month is chosen to update the contact networks and model the transmission events; keeping a very small (0.001 month) internal simulation timestep for other transitions in the model (e.g., population dynamics including birth and death). While the definition and duration of simulation time-steps is different across the literature and can range between hours (Chao et al. 2010) to years (Mellor et al. 2011) in different epidemic models, the appropriate choice depends on the scope and goal of study, and moreover, the extent of available data for parameter estimation in a smaller time scale. Given the scarcity of available data on daily contact patterns associated with the airborne transmission of TB (in contrast to the droplet-based transmission of most other respiratory diseases such as influenza) and the slow dynamics of TB (occurring on the order of years rather than weeks or months), we selected a timestep of one month. This timestep is long enough to facilitate crude estimation of monthly social-interaction patterns (described next), yet short enough to reveal the dynamics of TB transmission across a year-long period.

At the end of each month, the contact networks are updated (i.e., casual contacts of each infectious individual are selected randomly from the population), contact events are executed, and a probability of effective TB transmission through each contact is applied.
Consequently, the TB status of each individual is updated (e.g., effective transmissions cause a transition from “uninfected” to “early latent TB”), and the model proceeds to the next time step. Outputs are gathered at the end of each year, and averaged across many replications of the simulation model. In this section, we describe the population structure, contact network formulation, and the calibration procedure used in the epidemic simulation model.

1.1 Population Structure

We assume that, at baseline, household size follows a discrete triangular distribution, ranging from 1 to 10 with a mode of 5 people per family, which corresponds to low socioeconomic status typical of the moderate incidence TB settings in India and Brazil (System 2013). The natural mortality rate is defined at the individual level, and corresponds to a life expectancy of 73 years (Google 2010). The natural birth rate for new individuals in each household at each month is defined as a decreasing function of the family size and the natural mortality rate, such that the average population size remains constant, and the initial household-size distribution keeps a stable shape through the course of the simulation period (Figure 1). This function is defined with regard to an empirical function of family sizes and the individual’s natural death rate; the coefficients are tuned by trial and error.
For purposes of simplicity in this model, we do not incorporate the age structure, as pediatric TB does not contribute substantially to TB transmission, and demographic changes are unlikely to alter the impact of contact tracing on TB incidence over a short time period of 5-10 years. Since children (especially infants) may develop rapidly progressive TB and represent an important focus of household contact tracing efforts (Noertjojo et al. 2002; Clark & Cant 1996), our model may appropriately estimate effects on TB incidence but underestimate the mortality benefit of contact tracing for pediatric TB. Moreover, we also do not explicitly incorporate HIV, as empirical data on the natural history of untreated HIV-associated TB are sparse and conflicting.

1.2 Natural History of TB

TB natural history is modeled at the individual level using five main TB health states (Figure 2). A person is assumed to be born in full health and susceptible to TB. Upon successful transmission of the disease, the person enters the early latent TB (ELTB) state for a period of five years. During this time, individuals are at high risk of progression to active TB; this probability declines with each year in the ELTB state (i.e., the ELTB state consists
of five sub-states, each lasting for one year), according to estimates of adult TB progression following infection (Vynnycky & Fine 1997). After five years post-infection, individuals enter the late latent TB (LLTB) state, during which time they have a very low but non-zero probability of progression to active disease in each year. Reinfection can occur during the early or late latency period, and is modeled as a return to the first year of the ELTB state. Individuals with latent TB are assumed to have a degree of immunologic protection from reinfection (Andrews et al. 2012; Sutherland et al. 1982), modeled as a reduced risk of infection, compared with a susceptible individual, upon contact with an infectious individual.

![Diagram of TB states and transitions]

Figure 2: Agent-Based Model Outline. The individuals are subject to the natural mortality rate (not shown) at all times. The parameter definitions and calibrated values are provided in Table 1 of the manuscript.

Patients with active TB (ATB) are infectious and subject to an increased probability of mortality. Infectiousness is modeled as zero at the start of the active period, increasing thereafter as a linear function of time for the first nine months of disease (i.e., as the
bacillary burden grows), and stable thereafter at the maximum level until the individual is diagnosed and treated, or dies. The maximum transmission probability is calibrated to provide the specified incidence rate at baseline, thereby reflecting a combination of intrinsic transmissibility and external factors (e.g., crowding, poverty) that may modify the probability of effective TB transmission. Upon diagnosis and initiation of effective treatment, or alternatively through spontaneous resolution, individuals with active TB enter the non-infectious recovered (REC) state. We model the spontaneous recovery rate based on a systematic review of TB natural history (Tiemersma et al. 2011) and calibrate the treatment rate such that individuals remain infectious for an average of 11 months prior to treatment, the global mean as estimated by the WHO (World Health Organization 2013b) (i.e., estimating the disease duration through the ratio of prevalence to incidence, and reducing it by 6 months to account for the average duration of treatment). These individuals are also subject to a risk of reinfection, as described above.

1.3 Contact Network Formulation

We define $S=\{SUS, ELTB, LLTB, ATB, REC\}$ as a set of possible TB-related health states, where $s^{\text{tin}}$, and $s^{\text{tout}}$ denote the time for entering to, and leaving from state $s$, $s \in S$. The model follows a population of constant average size of $P$ through time. We let $p, p = 1, \ldots, P$ denote the $p$th individual in our model, and let $p.s(t)$ denote person $p$’s health state at time $t$ throughout the simulation period.

Assuming that only individuals with active TB are infectious, we let $p.\text{Inf}(t)$ denote the person $p$’s infectiousness at time $t$, where $p.s(t) = ATB$. We define the infectiousness value as a linear function of time from the onset of the disease ($p.s^{\text{tin}}$) as
where the infectiousness increases from 0 to the value of $I_{max}$ during a disease period of $I_{duration}$ (months), and stays at the maximum level for the rest of infectiousness period. Moreover, we assume that latently infected individuals develop partial immunity toward reinfection of TB, and note this immunity with $p.\text{Imm}(t)$, where $p.s(t)=ELTB$ or $p.s(t)=LLTB$.

We model two sources of disease transmission, among household members (close contacts) and community members (casual contacts). Each contact type is associated with an effectiveness parameter ($Eff_{HH}$ and $Eff_{COM}$) that reflects the closeness of the contact between the two individuals, and ultimately the likelihood of disease transmission upon an infectious contact through each network.

The simulation model runs in discrete time steps representing months. In order to use computational power efficiently, we model only the contact events for the infectious individuals at each time step. Let $p$ denote the only single infectious person in the population $P$ at time $t$. Person $p$ contacts with his/her household members and $C$ randomly chosen members of the community, where $C$ is the average frequency of casual contacts over a month. With these definitions, the probability of disease transmission ($P_{trans}$) from the infectious person $p$, to a non-infectious person $q$ at the end of first time-step (one month) can be modeled as follows

$$P_{trans}(t) = Pr_{HH-contact}(p,q,t) \times \left[ p.\text{Inf}(t) \times (1- q.\text{Imm}(t)) \times Eff_{HH} \right] + Pr_{Com-contact}(p,q,t) \times \left[ p.\text{Inf}(t) \times (1- q.\text{Imm}(t)) \times Eff_{Com} \right]$$
where $Pr_{HH-contact}(p,q,t)$ denotes the probability of a close contact between individual $p$ and $q$ at time $t$ (equal to 1 if $p$ and $q$ are household members and 0 otherwise), and $Pr_{Com-contact}(p,q,t)$ denotes the probability of a random casual contact between individual $p$ and $q$ at time $t$, which can be estimated as $\frac{C}{P(t)}$.

Due to uncertainty of the contact network structure, we considered two scenarios representing settings of high (household-driven) and low (community-driven) household transmission with an incidence rate representative of the current global mean (120 per 100,000 yr$^{-1}$). Table 1 shows a list of calibrated contact network parameters under each scenario. Assuming fixed values of $I_{max}$ and $I_{duration}$ for our baseline analysis (the values were later changed for the sensitivity analysis), we estimate the average number of casual contacts ($C$) from previous studies on transmission pattern of airborne disease and TB (Mossong et al. 2008). Moreover, we calibrate the contact’s effectiveness coefficients ($Eff_{HH}$ and $Eff_{Com}$) to provide the specified household transmission ratio in each scenario; i.e. in order to provide a proportion of 22% of total incidence through household transmissions in the community-driven scenario, the values of $Eff_{HH}$ and $Eff_{Com}$ are tuned to 1 and 0.035 accordingly.
### Table 1: Simulation input parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease delay to reach maximum infectiousness ($I_{duration}$)</td>
<td>9 months</td>
</tr>
<tr>
<td>Maximum infectiousness ($I_{max}$)</td>
<td>0.1</td>
</tr>
<tr>
<td>Number of casual contacts per month ($c$)</td>
<td>300 people</td>
</tr>
<tr>
<td>Effectiveness of household contact ($Eff_{HH}$) in (Community-driven, household-driven) scenario</td>
<td>(1, 4.5)</td>
</tr>
<tr>
<td>Effectiveness of casual contact ($Eff_{com}$) in (Community-driven, household-driven) scenario</td>
<td>(0.035, 0.0292)</td>
</tr>
</tbody>
</table>

**A Note Regarding the Household Transmission Ratio**

Previous studies on the transmission dynamics of TB in households use definitions for the household transmission ratio, and apply different computation methods to estimate its value (Classen et al. 1999; Verver et al. 2004; Madico et al. 1995). These methods are selected with regard to the type of available transmission data and design of the study. Here, we adopt a generally accepted definition of *household-transmission-ratio* (HHTR) as “the proportion of people with active TB who were infected by a household member” (Verver et al. 2004). We estimate this value by computing the average proportion of incident active TB for which the corresponding infection (assuming that the most recent "successful" infection is the one that leads to disease) originating through a household contact. This definition, however, does not incorporate the poorly understood dynamics of mixed-strain infections, and it also does not account for the fact that some infections may not lead to active disease (e.g., individuals who are reinfected with a second strain and whose first strain subsequently reactivates). In other words, we define the "successful infection" as one that causes the new strain to become the dominant strain in an individual, without considering mixed infections. We therefore define a second quantity, namely the total-household-transmission-ratio (T-HHTR), as the proportion of total transmission events (including new infection of
susceptible or reinfection of previously infected individuals) that occur among household members. Table 2 shows the estimated value for each output and under each scenario in our model. Accounting for all new infections and re-infection events among household members, T-HHTR > HHTR because household contacts of individuals with active TB are more likely to experience reinfection events than are general members of the community.

Table 2: The household transmission ratios. The values represent [Average (95% confidence interval)].

<table>
<thead>
<tr>
<th>Output</th>
<th>Community Driven</th>
<th>Household Driven</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHTR</td>
<td>0.19 (0.111, 0.262)</td>
<td>0.32 (0.231, 0.412)</td>
</tr>
<tr>
<td>T-HHTR</td>
<td>0.22 (0.14, 0.3)</td>
<td>0.5 (0.399, 0.593)</td>
</tr>
<tr>
<td>Proportion of household</td>
<td>0.26 (0.25, 0.27)</td>
<td>0.5 (0.49, 0.51)</td>
</tr>
<tr>
<td>members infected by a case</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In calibrating each scenario, we consider the most recent "successful" infection as being the one that leads to disease among individuals with incident active TB. When we include reinfection in our simulation model, a HHTR (e.g., 19%) proportion of infections leading to active disease corresponds to a T_HHTR (e.g., 22%) of all transmission events (initial infection or reinfection) occurring within the household.

1.4 Simulation Calibration Procedure

In order to generate the initial simulated population, we created an equilibrium state using a long model run (1000 years), and determined the initial proportion of population in each health state. Furthermore, we added an additional transient burn-in period of 100 years for each simulation, and start our analysis from year 101. This helped to eliminate the early bias of main outcomes due to simulation initialization, and achieved a relative steady state for additional perturbation (e.g., intervention). For each experiment, we replicated the
simulation model for more than 1000 runs; as stated in the main text, this number was selected as providing a relative precision of at worst 1% for the incidence rate (i.e., incidence rate falls between 120 ± 1.2 per 100,000/year). In this section, we provide additional details on our calibration procedure of the TB epidemic simulation model.

1.4.1 Determining the Initial Population Setting

The simulation model initiates from an existing TB endemic, and follows the course of disease epidemiology through time. The initial conditions are determined from the equilibrium state of the simulation model calibrated to the baseline incidence level. For this purpose, we initialize the calibrated model from a random initial point, and let the model run to reach its equilibrium state over >1000 years. Consequently, the equilibrium proportion of susceptible, early latent, late latent, infected and recovered individuals is estimated as {0.57, 0.033, 0.36, 0.0012, 0.034}, and is used as the initial endemic state of the simulation model.

1.4.2 Computing Simulation Transient Time

We follow a similar approach to Kasaie et al. (2013) in order to determine the simulation transient-time period (Kasaie et al. 2013). Figure 2 shows the main ABS outputs over 200 years. The external infection ratio represents the proportion of disease incidence due to reactivation or progression among individuals who were infected at time zero, as a part of the existing endemic (Panel A). Since such infections did not originate within the simulation period, they cannot be used for the analysis of transmission. Moreover, due to the stochastic nature of the simulation model and disease dynamics, the annual incidence rate shows a non-stationary pattern during the first decades (Panel B). We choose the simulation transient period such that the external infection ratio is less than 5%, and the incidence rate has a stationary mean (Figure 2).
We test the external infection ratio mean using a one-tailed t-test, and test the stationarity using multiple comparison tests of the incidence mean in consecutive 5-year periods after the transient time (using Bonferroni-adjusted paired t-tests). Consequently, we choose a transient period of 100 years for the simulation model (burned in at the beginning of each simulation run), and start our analysis of the output in year 101.

1.4.3 Computing the Required Number of Simulation Replications

We estimate the required number of runs using the maximum of two approximation techniques as

Figure 2: Simulation output pattern. A. Annual external infection ratio. B. Annual incidence rate per 100,000 person year.
where \( n_0 \) is the initial number of replications, and \( s_0 \) and \( h_0 \) are the standard deviation and desired 95% half-width interval around the average response value (Law & Kelton 2000). From these initial \( n_0 \) runs, we choose the maximum required number to provide the desired relative precision, \( e = \frac{h}{\bar{y}} \), for the simulation output \( Y \), and selected the maximum of the two suggested values. Choosing the annual incidence level as the main output of interest, we estimate a minimum of \( n=1020 \) replications for the initial experiments to provide a relative precision of 0.01 around the average response. In the later steps of analysis, therefore, we use a minimum of 1020 simulation replications to ensure the desired relative precision of our results.

2 Results

Table 3 provides a summary of model outputs under the community-driven and household-driven scenarios at the equilibrium state.
Table 3: Simulation Output Summary. Values represent [Average (95% Confidence Interval)]. All values are in the in the original scale of the simulated population of 10200 individuals.

<table>
<thead>
<tr>
<th>Output</th>
<th>Community-Driven</th>
<th>Household-Driven</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean duration of infectiousness (months)</td>
<td>10.94 (10.93, 10.96)</td>
<td>10.95 (10.94, 10.96)</td>
</tr>
<tr>
<td><strong>Number of individuals developing active TB per year</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fast Progression, year 1 post-infection</td>
<td>6.36 (6.31, 6.4)</td>
<td>6.5 (6.46, 6.54)</td>
</tr>
<tr>
<td>Fast Progression, year 2 post-infection</td>
<td>2.41 (2.4, 2.43)</td>
<td>2.47 (2.46, 2.49)</td>
</tr>
<tr>
<td>Fast Progression, year 3 post-infection</td>
<td>0.74 (0.73, 0.74)</td>
<td>0.76 (0.76, 0.77)</td>
</tr>
<tr>
<td>Fast Progression, year 4 post-infection</td>
<td>0.48 (0.47, 0.48)</td>
<td>0.49 (0.48, 0.49)</td>
</tr>
<tr>
<td>Fast Progression, year 5 post-infection</td>
<td>0.15 (0.15, 0.15)</td>
<td>0.15 (0.15, 0.16)</td>
</tr>
<tr>
<td>Fast progression, cumulative</td>
<td>10.14 (10.07, 10.2)</td>
<td>10.38 (10.32, 10.44)</td>
</tr>
<tr>
<td>Slow progression (&gt;5 yrs post-infection)</td>
<td>1.86 (1.85, 1.87)</td>
<td>1.83 (1.82, 1.84)</td>
</tr>
<tr>
<td>Total incidence (fast + slow progression)</td>
<td>11.99 (11.92, 12.06)</td>
<td>12.21 (12.14, 12.28)</td>
</tr>
<tr>
<td><strong>Number of individuals removed from active TB per year</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mortality</td>
<td>1.31 (1.31, 1.32)</td>
<td>1.34 (1.33, 1.35)</td>
</tr>
<tr>
<td>Recovery</td>
<td>1.31 (1.31, 1.32)</td>
<td>1.34 (1.33, 1.35)</td>
</tr>
<tr>
<td>Treatment</td>
<td>9.22 (9.16, 9.27)</td>
<td>9.38 (9.32, 9.43)</td>
</tr>
<tr>
<td>Total</td>
<td>11.84</td>
<td>12.06</td>
</tr>
<tr>
<td><strong>Number of individuals in each compartment at steady state</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active TB</td>
<td>12.01 (11.85, 12.17)</td>
<td>12.31 (12.14, 12.48)</td>
</tr>
<tr>
<td>Early latent TB</td>
<td>334.01 (329.54, 338.48)</td>
<td>345.06 (340.52, 349.6)</td>
</tr>
<tr>
<td>Late latent TB</td>
<td>3725.62 (3719.96, 3731.28)</td>
<td>3651.14 (3635.37, 3666.91)</td>
</tr>
<tr>
<td>Recovered</td>
<td>392.85 (391.93, 393.77)</td>
<td>353.74 (352.02, 355.46)</td>
</tr>
</tbody>
</table>
2.1 Analysis of the Timing of Infections

2.1.1 Timing of New Infections

When we calibrated the treatment rate to provide a mean 11-month duration from onset of infectiousness to resolution (by death, spontaneous cure, or initiation of treatment), over 50% of infections to casual contacts nevertheless originated from individuals who had been infectious for more than 11 months (Figure 3-B, blue line). Diagnosis and contact tracing occurred at a mean 10.8 months (household-driven) to 11.5 months (community-driven) after onset of infectiousness, by which time 75% (community-driven) to 95% (household-driven) of new household infections had already occurred. The monthly rate of infecting casual contacts peaked at the point of attaining maximum infectiousness (9 months in our primary model), and did not depend on the intensity of household transmission; by contrast, the rate of household infections peaked three to five months sooner, with more intense household infection resulting in faster “saturation” of household contacts (Figure 3-A).

![Figure 3](image-url)

**Figure 3**: Timing of secondary new infections from a single case, assuming an infectious period of mean 11 months.
Panel A shows the number of new infections (i.e., transition from susceptible to early latent TB, similar to a tuberculin skin test conversion, excluding reinfections) per month occurring from an average case to household contacts (red) and casual contacts (blue); panel B shows the proportion of all new infections that occur by a given month in the infectious period, assuming an average duration of infectiousness of 11 months.

2.1.2 Timing of Total Infections

Figure 4 shows the estimated timing of infections among close and casual contacts in each scenario. This estimate is based on all of the transmission events in our model, including both new infections of previously uninfected individuals and reinfections of previously infected individuals. The trajectory of cumulative infections through close contacts is very similar in both scenarios (panel B) despite wide variation in the proportion of infections occurring in the household. This finding suggests that the “median infection” (time at which 50% of all infections are completed) occurs about 3 months earlier for household transmission than for casual transmission, regardless of the intensity of household transmission.
This Figure corresponds to Figure 3, only showing all secondary infections rather than new/initial infections only. Panel A shows the number of total infections per month (i.e. tracking all infections including reinfections), and panel B shows the cumulative proportion of total infections. The solid lines represent the “community-driven” scenario (78% of all TB transmission occurs in the community) and the dotted lines the “household-driven” scenario (50% of all TB transmission occurs in the community). Red lines denote close (household) contacts and blue lines denote casual contacts.

2.1.3 Sensitivity Analysis of Timing of Infection to Variation in the Infectiousness-Versus-Time Relationship

The timing of the “median infection” depends on the shape of the infectiousness-versus-time curve. To explore this further, we considered alternative scenarios in which the infectiousness of active TB increased linearly from zero to a maximum over six months, nine months (the basecase scenario), and 12 months. Figure 5 shows the timing of new infections in each of these three scenarios. Notably, the timing of the “median infection” for
both household and casual contacts did not vary widely with different shapes of the infectiousness-versus-time trajectory.

2.1.4 One-Way Sensitivity Analyses: Timing of Infection

We subsequently evaluated the sensitivity of TB infection timing (measured as the time of “median infection” in both household and casual contacts) to variation of model input parameters, as shown in Table 1 of the manuscript. For the purpose of comparability, we calibrated each experiment to the baseline incidence level of 120 per 100,000/y and a household transmission ratio of 22%, reflective of our community-driven scenario (except the scenarios in which incidence was varied). In experiments describing changes in incidence, we assumed that these changes in incidence reflected only casual/community transmission; in other words, we assumed that the per-month frequency and effectiveness of household transmission does not vary with the underlying incidence in the community.

Figure 5: Sensitivity analysis of timing of new infections under different assumptions regarding the duration of disease before reaching maximum infectiousness. A. Number of new infections per month. B. Cumulative proportion of new infections. The solid lines represent close contacts and the dotted lines casual contacts.
Thus, we started from our baseline model (community-driven scenario, incidence of 120 per 100,000/year), and tuned the incidence level by changing only the community contact rate \((\text{Eff}_{\text{Com}})\). As a result, the household transmission ratio is higher in low-incidence settings. Specifically, at incidence rates of 8, 120, and 300 per 100,000/yr, the corresponding household transmission ratios are 39%, 22%, and 14%, respectively.

For the variation of fast and slow progression rates, we note an intrinsic dependency between these two parameters to maintain a fixed incidence level – for a given incidence and fast progression rate, the slow progression rate is defined. Thus, we defined the experiments varying fast and slow progression according to the proportion of the incidence due to recent infection (i.e., fast progression from early latent TB), changing this value from 30% (corresponding to 5% fast progression cumulative over 5 years, and 0.23%/year slow progression rate) to 96% (corresponding to 17% cumulative fast progression over 5 years, and 0.01%/year slow progression rate). Table 4 shows the results of sensitivity analysis.

Table 4: Sensitivity analysis of the timing of infections. Low and high parameter values are set to specified levels in Table 1 of the manuscript.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Average time of median infection (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low/Close</td>
</tr>
<tr>
<td>Household size mean (range)</td>
<td>9.1</td>
</tr>
<tr>
<td>Life expectancy</td>
<td>9.0</td>
</tr>
<tr>
<td>ATB mortality rate</td>
<td>9.7</td>
</tr>
<tr>
<td>Early latency duration</td>
<td>9.0</td>
</tr>
<tr>
<td>Proportion of TB due to recent infection</td>
<td>9.1</td>
</tr>
<tr>
<td>Recovery rate</td>
<td>9.8</td>
</tr>
<tr>
<td>Diagnosis and Treatment rate</td>
<td>12.0</td>
</tr>
<tr>
<td>Latent immunity toward re-infection</td>
<td>11.3</td>
</tr>
<tr>
<td>Incidence rate</td>
<td>8.6</td>
</tr>
<tr>
<td>Disease duration to reach maximum infectiousness</td>
<td>8.0</td>
</tr>
</tbody>
</table>
Thus, in none of our one-way sensitivity analyses did the “median infection” among households vary outside the range of 7.9-11.3 months, nor did the “median infection” among casual contacts vary outside the range of 10.1-13.0 months. The one exception to this was the treatment rate, which served as a proxy for the duration of infectiousness. When the treatment rate was increased to 2.0/year (i.e., the duration of disease was 5.4 months until treatment, mortality, or self-cure), the median household infection occurred at 5.7 months, and in the community at 7.4 months. Similarly, when the treatment rate was slowed to 0.5/year (i.e., a 16-month duration of the disease), the median household infection occurred at 12.0 months, and in the community at 15.2 months. In other words, under all explored parameter variations, the median household infection consistently occurred 1-4 months before the mean duration of infectiousness was reached, and 1.5-3 months before the median infection occurred among casual contacts.

2.2 Effectiveness of Case-Finding Interventions

As described in the paper, we considered a household contact tracing (HHCT) intervention, in which a proportion of passively diagnosed TB cases (“index cases”) are selected, and their household members are then screened and treated for active TB. We extended this intervention to include preventive therapy (PT) that is capable of reducing the future risk of reactivation (though not reinfection) by 70%. Moreover, we considered a community-based active case finding (ACF) strategy, in which a randomly-selected set of households is chosen at annual intervals, screened, and treated for TB. The number of households in this program was calibrated to provide a similar incidence reduction (10%) as the HHCT strategy in the community-driven scenario – which required a total of 700 randomly selected
households in a population of 10,200 individuals over 5 years. Table 2 shows the programmatic indicators of the case-finding strategies.

Table 5: Programmatic indicators of case-finding interventions. Values represent [Average (95% half-width confidence interval)]

<table>
<thead>
<tr>
<th>Community-Driven</th>
<th>Total Number of People Screened</th>
<th>% of Sample Found to be:</th>
<th>Average Duration of Infectiousness Before Screening (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHCT</td>
<td>199.2 (3.1)</td>
<td>31.6</td>
<td>31.2</td>
</tr>
<tr>
<td>HHCT+PT</td>
<td>189 (3)</td>
<td>31.1</td>
<td>30.8</td>
</tr>
<tr>
<td>ACF</td>
<td>3575.6 (1.6)</td>
<td>3.2</td>
<td>36.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Household-Driven</th>
<th>Total Number of People Screened</th>
<th>% of Sample Found to be:</th>
<th>Average Duration of Infectiousness Before Screening (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHCT</td>
<td>195.9 (3)</td>
<td>57.1</td>
<td>21</td>
</tr>
<tr>
<td>HHCT+PT</td>
<td>179.2 (2.8)</td>
<td>56.4</td>
<td>20.8</td>
</tr>
<tr>
<td>ACF</td>
<td>3574.5 (1.6)</td>
<td>3.3</td>
<td>35.8</td>
</tr>
</tbody>
</table>

2.3 Sensitivity Analysis of Household Contact Tracing Effectiveness

For all interventions, the transmission impact was a nearly linear function of the diagnostic test sensitivity (Figure 6-A), and the proportion of screened TB cases who were effectively diagnosed and treated (Figure 6-B). Thus, for example, using a screening algorithm that only identified and treated 45% of eligible TB cases (e.g., through any combination of imperfect test sensitivity and losses to follow-up before treatment) reduced TB incidence by only half as much as an algorithm that diagnosed and treated 90% of eligible cases.
Sensitivity analysis of the transmission impact of contact tracing (Figure 6 of the manuscript) suggested a close dependence to the underlying incidence rate, having reduced impact in both low-incidence settings (where the majority of incident TB is due to reactivation) and very high-incidence settings (where the majority of incidence is due to community transmission). Figure 7 shows the transmission impact of the HHCT strategy over a wide range of incidence rates, varying incidence purely as a function of casual transmission (i.e., same household transmission effectiveness regardless of background incidence rate), as above. The effectiveness of HHCT increases in an approximately linear fashion until the underlying TB incidence reaches a level of 50 per 100,000/year, defining settings in which 80% or more of incident TB is due to recent transmission (Figure 8, orange dotted line). Beyond this point, the transmission impact begins to decrease somewhat with increasing background incidence, reflecting a declining ratio of household transmission (Figure 8, green dotted line). The results imply that the population-level impact of

2.4 Sensitivity Analysis of Household Contact Tracing Effectiveness to Variation of Underlying TB Incidence

Sensitivity analysis of the transmission impact of contact tracing (Figure 6 of the manuscript) suggested a close dependence to the underlying incidence rate, having reduced impact in both low-incidence settings (where the majority of incident TB is due to reactivation) and very high-incidence settings (where the majority of incidence is due to community transmission). Figure 7 shows the transmission impact of the HHCT strategy over a wide range of incidence rates, varying incidence purely as a function of casual transmission (i.e., same household transmission effectiveness regardless of background incidence rate), as above. The effectiveness of HHCT increases in an approximately linear fashion until the underlying TB incidence reaches a level of 50 per 100,000/year, defining settings in which 80% or more of incident TB is due to recent transmission (Figure 8, orange dotted line). Beyond this point, the transmission impact begins to decrease somewhat with increasing background incidence, reflecting a declining ratio of household transmission (Figure 8, green dotted line). The results imply that the population-level impact of
household contact tracing is likely to be highest in medium-incidence settings, where most
disease is due to recent infection, but community transmission rates are not so high as to
overwhelm the contribution of household transmission.

Figure 7: Transmission impact of HHCT in settings with different TB incidence.

Figure 8: Proportion of household transmission and recent infections in settings with different TB incidence.
3 References


Chao, D., 2013. Calibrating an influenza epidemic model.


Google, 2010. Public Data: Life Expectancy. Available at: https://www.google.com/publicdata/explore?ds=d5bneppjoof8f9&met_y=sp_dyn_le00_in&dim=country:USA&dl=en&hl=en&q=average life expectancy#ctype=l&strail=false&bcs=d&nselm=h&met_y=sp_dyn_le00_in&scale_y=lin&ind_y=false&rdim=region&dim=region:MNA:SSA:SA.


Chapter 3: Estimation Bias of the Ratio of Tuberculosis Incidence Due to Recent Transmission Using DNA Fingerprinting Data
Abstract

Analysis of population-based DNA data continues to serve as the main method to estimate the proportion of TB incidence due to recent transmission, which in turn has important implications for understanding the dynamics of transmission and policymaking. Previous studies have identified a number of factors affecting the precision of this approach in various settings, but the exact relationship of such factors remains uncertain. In this study, we aim to quantify the role of such factors on estimation bias of the proportion of incidence due to recent transmission. Using an agent-based simulation model of TB as a virtual laboratory, we implement a sequence of statistically controlled experiments with regard to combinations of several factors. The results enable us to compute the estimation bias for various levels of each factor, and ultimately provide a decision-support tool for adjusting the estimation error in a variety of settings.
1 Introduction

Upon infection with mycobacterium tuberculosis, individuals enter an early latency state that is associated with a high risk of progression to active TB (*primary infection*). This period lasts for five years, and the probability of progression declines with each year in the early latent state, according to estimates of adult TB progression following infection (Vynnycky & Fine 1997). After five years post-infection, individuals enter the late latency state, during which they have a very low but non-zero probability of progression to active disease in each year (*reactivation*). The latently infected individuals are subject to risk of reinfection with the disease, but the previous TB infections have been shown to induce certain levels of immunity toward reinfection during the latency period (Andrews et al. 2012). Given this, any new case of TB can be the outcome of a recent transmission event associated with primary infection, or the result of the reactivation of a latent infection acquired some years previously.

The proportion of incidence due to primary infection is a close estimate of the rate of TB recent transmission, and has important implications for understanding the transmission dynamics and designing control strategies. For example, the majority of TB incidents in developed countries such as the US and Canada are expected to be a result of disease reactivation associated with previous transmission events many years ago when the incidence of TB was considerably higher. Eventually this reservoir of latent TB will be exhausted and then, in theory, more primary cases can be expected, which will arise mainly due to immigration or infections acquired outside of these countries. TB-controlling strategies in such circumstances, therefore, should focus more heavily on screening programs for immigrants and prevention strategies for the rest of the population. By contrast, in high-incidence countries such as India and South Africa, on-going transmission is a problem that requires identification of more specific performance targets for TB.
control. While TB-control goals include improvements in case detection, as well as reductions in diagnostic delay and time to effective treatment, the implications of the relative levels of primary TB for assessing the effectiveness of current control measures is of great importance.

With lack of TB diagnostic techniques to differentiate the original timing of transmission, questions regarding the proportion of incidence due to primary infection versus reactivation remained unanswered until 1990 when molecular fingerprinting was introduced. Molecular subtyping (e.g., by use of the international standardized method of restriction fragment–length polymorphism) allows characterization of specific strains of *Mycobacterium tuberculosis* (MTB) on the basis of their DNA pattern (Vynnycky et al. 2001). Comparison of fingerprints from patients’ isolates during TB outbreaks has demonstrated matching patterns among those who were infected from a common source. This suggests that patients with the same MTB DNA pattern constitute an epidemiologically linked cluster (Takahashi et al. 1993; Chevrel-Dellagi et al. 1993). Given the short period of TB development among these patients, the clustering level (i.e., the proportion of TB cases that share identical DNA patterns) can be used as a general identifier of the extent of primary infections attributable to recent TB transmission (Small et al. 1994; Chevrel-Dellagi et al. 1993). For example, 57% of all patients with tuberculosis in Denmark with onset of disease during 1992–1995 were infected with strains of TB that were part of a cluster, and the disease in most of these patients has been attributed to recent transmission (Yang et al. 1995).

Though it is recognized that the extent of clustering of isolates from tuberculosis cases in a given population is related to the amount of disease attributable to recent transmission, the relationship between the two statistics is poorly understood. Previous studies have identified underlying factors affecting the estimates of the clustering level (Murray 2002a;
Glynn, Vyonycky, et al. 1999), and the relationship of clustering and recent transmission (Vynnycky et al. 2001; Murray 2002b). These studies suggest various sources of bias with regard to “sampling incompleteness”, as well as several disease- and host-related factors determining the “underlying cluster-size distribution”. Sampling incompleteness refers to the study’s ability to ascertain DNA fingerprint of members of the population (e.g., higher incompleteness correspond to missing more cases), which can be considered as a function of several factors including the sampling duration, study coverage, sample characteristics determining the individuals’ ability to provide valid DNA fingerprint, etc. The cluster size distribution, on other hand, refers to the underlying distribution of TB strains types among patients (i.e., frequency of patients with unique or similar TB strains).

Previous findings imply that estimates of recent transmission obtained by molecular methods cannot be compared across studies that have used different sampling fractions of population and in which the distribution of cluster size may vary. In such a setting, our aim is to develop a unifying framework for quantifying the role of various factors in the estimation bias of the ratio of TB incidence due to recent transmission (also referred to as “recent transmission rate” or similarly “primary infection rate”), that could be ultimately used as a decision-support tool in future epidemiological studies of TB.

We start our analysis by proposing a mathematical model of clustering-level estimation, which provides insight into the general behavior of clustering estimators in the presence of different cluster-size distributions and sampling-study coverage. We then extend our analysis to propose a simulation-based approach for studying the role of various factors on the estimation of bias with respect to the recent transmission rate. We use an agent-based simulation model of TB infection to model the dynamics of strain clustering and survival, and to generate various cluster-size distributions through time. We consider two factors in relation to the level of clustering, including the disease incidence level and the ratio of total
circulating strains in the population. We narrow the scope of this research to a setting of medium-to-high TB incidence, and levels of each factors in correspondence to available epidemic and demographic data form the representative countries (such as Brazil and India). Moreover, we define the sampling incompleteness with regard to the levels of study duration and sampling coverage. Using an agent-based simulation model of TB as a virtual laboratory, we implement a sequence of statistically controlled experiments with regard to combinations of each factor. The results enable us to compute the estimation bias for various levels of each factor, and can serve as a decision-support tool for adjusting the estimate of recent transmission rates.

2 Background

A major goal of population-based studies of TB molecular epidemiology is to differentiate and estimate the levels of incidence due to remote infection (reactivation of an old infection after years of latency) versus recent infection (progression to active disease in the early years of latency). In such studies, cases of tuberculosis are recruited from a defined area over an extended period, and isolates of *M. tuberculosis* (MTB) are collected from patients. The cultured isolates are then fingerprinted (using IS6110 or some set of markers (Murray & Nardell 2002)), and are analyzed for a clustering effect (i.e., those sharing a fingerprint with another isolate in the study sample are classified in clusters). The resulting clusters can be interpreted as epidemiologically linked chains of recently transmitted disease, and unique isolates to be cases of reactivation of the disease resulting from “remote” tuberculosis infection in the past (Murray, Megan 2002). Such studies have also been used to estimate measures of association among risk factors for recent transmission of TB, by comparing the distribution of an exposure in clustered cases to that in un-clustered cases (Verver et al.
2004; Vynnycky et al. 2001; Small et al. 1994). The studies typically report the proportions of unique and clustered cases and the odds ratios for risk factors for clustering.

2.1 Sources of Bias in Molecular Epidemiological Studies of TB

Despite the extensive use of DNA fingerprinting and clustering methods to estimate the relative contribution of primary infection to disease incidence, questions remain about the validity of the underlying assumptions and the precision of suggested methods to estimate rates of recent transmission. One criticism concerns the underlying assumption regarding the epidemiological link among the clustered cases. Several studies have reported controversial results in terms of their ability to establish epidemiological links between DNA clustered cases (Bishai et al. 1998; Small et al. 1994; van Deutekom et al. 2004).

This observation, however, can be partly due to studies’ limitations in establishing all ‘possible’ connections among people, which can as well reflect the importance of casual and random (community) contacts. Previous studies have contrasted the dynamics of close (e.g., intimate and prolonged contacts among household members) and casual (e.g., random encounters among neighbors and community members) contacts (Kasaie et al. 2014; Classen et al. 1999; Raffalli et al. 1996): many people exposed to a small risk may account for more cases than a few people exposed to high risk of transmission. Despite important contributions of casual contacts in some empirical settings (Raffalli et al. 1996; Cauchemez et al. 2011; Classen et al. 1999), however, epidemiological links for establishing such contacts are not always available. Moreover, clustering effects between individuals with no recent contact or epidemiological links may be due to some distant transmission link in the past, and reactivation of old infections in the present. The fraction of such events would depend on the historical incidence of the disease, and number of circulating strains in the population at that time. For example, the presence of two large historical DNA clusters early
in 1990 in Denmark is likely to result in clustering of reactivated cases in the future (Bauer et al. 1998).

Under the assumption of an epidemiological link between DNA clustered cases, it is reasonable to accept that these cases are part of the same chain of transmission. Further, in order to attribute the level of clustering to the rates of recent infection, one should consider the stability of DNA fingerprinting through time. The speed at which DNA fingerprint patterns change over time is an important factor in determining the extent to which the clustering observed in a population is associated with recent infection. The rates of changes of the molecular markers have been loosely described in terms of the molecular clock by analogy with those used to describe the evolution of proteins. In the case of a fast molecular clock, cases are only separated by very short serial time intervals and are more likely to share the same DNA fingerprint pattern (and hence to be ‘clustered’). A slow molecular clock, on the other hand, implies that even cases involved in distant chains of transmission (many years ago) may still appear clustered (Vynnycky et al. 2001; Glynn, Bauer, et al. 1999). Several studies have investigated the stability of DNA fingerprints, and suggested a half-life of 3.2 years for changes. The suggested stability in such studies, however, may not be the same as those seen in person-to-person transmission.

Given such phenomena, the interpretation of clustering observed in a population as implying recent transmission will further depend on the study duration and the age of the cases. Previous studies suggest that increasing the study duration (time window during which isolates were obtained from cases) is associated with an increase in the proportion of clustered cases (Glynn, Bauer, et al. 1999; Vynnycky et al. 2001; Murray, Megan 2002). Moreover, the observed clustering level among cases should be interpreted with regard to the population age-structure. One study shows that the proportion of TB disease attributable to recent transmissions is higher among young individuals than for the elderly who have had
many years to become infected and have a higher chance of reactivation (Vynnycky et al. 2001). Their analyses therefore imply that the clustering in a given time interval is likely to underestimate recent transmission for younger individuals, and overestimate that for older individuals.

Moreover, the criteria for making precise estimates of underlying clustering levels requires representative samples of population that are free from both major random and systematic errors (Glynn, Vyonycky, et al. 1999). While population-based molecular studies are often based on random or convenience samples drawn from available clinical isolates of TB, the size and inclusion criteria of DNA samples is another source of bias in estimating the clustering level. Small samples tend to miss more cases of TB through time, and underestimate the clustering effect. This, in turn, results in underestimation of the extent of recent transmission and overestimation of reactivation. The magnitude of the bias incurred by sampling strategies depends both on the sampling fraction and the frequency distribution of sizes of clusters in the population (Glynn, Bauer, et al. 1999).

Finally, various analytical methods for analyzing DNA samples can result in different estimates of the clustering level. The most widely used methods include the “n” and “n-1” methods. The “n” method uses the number of all cases falling into clusters as the estimator of clustered cases (in which n represents the cluster size). The “n-1” method assumes that one case per cluster is a source case (which may as well be due to reactivation), and thus removes one case per cluster from the counts of clustered cases (such that the number of secondary cases is one minus the cluster size n). Previous studies have contrasted the performance of the two methods in various settings, and have suggested that, despite the systematic bias included in both estimators, the “n-1” method is more appropriate to ascertain the number of people with primary versus reactivated disease (Murray, Megan 2002).
Figure 1 shows an outline of the estimation procedure for the proportion of TB incidence due to recent transmission starting from population sampling to DNA fingerprinting of sampled patients and analysis of DNA clusters, and finally interprets the rate of recent transmission from the estimated level of clustering. Moreover, the figure summarizes our discussion in terms of uncertainties and suspected sources of bias in each step of this procedure.

2.2 Previous Research

Despite the wide application of molecular epidemiologic methods (such as ‘n-1’ cluster analysis) in estimating the rate of TB recent transmission, and wide discussion regarding the
shortcomings of this approach, only a few studies have addressed the issue. A number of researchers have reviewed the objectives and design of molecular epidemiologic studies of TB, and consider the impact of different analytical approaches in estimating DNA clustering levels (Murray & Salomon 1998; Glynn, Bauer, et al. 1999; Murray & Nardell 2002). These studies suggest incomplete study coverage and underlying cluster-size distribution as main factors determining the estimation bias.

The true underlying cluster distributions cannot be known in the absence of complete sampling, and in turn depend on a variety of host- and epidemic-related factors. Murray (2002) explores social and demographic determinants of cluster distributions (Murray 2002a). Using an individual-based micro-simulation of TB transmission, researchers have considered the role of multiple factors on cluster-size distribution; including TB control strategies, HIV coinfection, prevalence of latent TB, and different population age structures. The results suggest that multiple host-related factors contribute to wide variation in cluster-size distributions, which in turn can have a strong impact on the precision of cluster-analysis methods in estimating levels of recent infection.

Glynn et al. (1999) consider the influence of incomplete sampling on estimates of clustering (Glynn, Vyonycky, et al. 1999). Researchers have used data from two populations with different reported cluster patterns (similar overall clustered proportions) as starting populations, and consider a series of artificial populations to study the effect of clustering on random samples of these populations. The hypothetical population includes 100 individuals and represents different underlying clustering patterns (clusters arranged as pairs, triplets (totaling 99 individuals), fours, or fives, giving a clustered proportion of 100 percent). The authors found a positive relationship between the sampling fraction and the (average) proportion found to be clustered, and suggest that the extent of the underestimation depends
on the cluster structure of the population: the influence of incomplete sampling is weaker in populations with large clusters than in those with small clusters.

Murray and Alland (2002) extend this analysis by applying analytical methods in addition to simulation (Murray, Megan 2002). Using a Monte-Carlo simulation model (Murray 2002a), they investigate five scenarios for various hypothetical cluster-size distributions, and study the estimation bias in relative proportion of clustered and un-clustered cases by different cluster-analysis methods (“n” vs. “n-1”). Their results suggest that, while both methods yield biased results after sampling, the “n-1” method is more appropriate to ascertain the number of people with primary versus reactivated disease.

Even under 100% certainty, interpretation of the clustering level attributable to recent infection may still be a matter of uncertainty. Vynnycky et al. (2000) consider the effect of age and study duration on the estimated level of clustering and the proportion of disease attributable to recent infection. They develop an age-structured model of transmission based on The Netherlands’s data between 1993 and 1997. Their results suggest that the observed level of clustering in a population depends on the age structure, and the clustering in a given time interval is likely to underestimate recent transmission for younger individuals, and overestimate that for older individuals.

An overview of previous studies suggests several sources of error in estimating the clustering level of a population, and interpreting the amount of primary infection on that basis. Moreover, the results imply that estimates of recent transmission obtained by such molecular methods cannot be compared across studies that have used different sampling fractions of the population, and in which the distribution of cluster size may vary. In such a setting, our objective is to develop a unifying framework for quantifying the role of various factors in estimation bias of TB recent infection ratio. We explore the range of potential error under a variety of experimental settings with regard to levels of various factors, and
develop a predictive model as a function of these factors. The suggested model can be used to estimate expected error under various settings, and adjust the estimations of the primary infection ratio.

3 Mathematical Analysis of Bias in Estimators of the Clustering Level

Performance of the clustering methods depends on sampling characteristics, and it’s been shown that the estimator often tends to underestimate the level of clustering. In this section, we discuss an analytical approach to describe the estimation bias of clustering due to incomplete sampling. Because of uncertainty regarding the underlying cluster-size distribution, previous studies have analyzed the results in the domain of certain scenarios (Glynn, Vyonycky, et al. 1999; Murray, Megan 2002). Here, we extend the analysis by considering a simple model of cluster-size distribution, and study the general characteristics of this model in various settings. While the suggested linear-approximation model may not be a realistic representation of the true cluster-size distribution, it enables us to model a wide class of distributions and provides insight into the general behavior of clustering estimators in the presence of different cluster-size distributions and sampling study coverage. The finding are used in our later analysis regarding estimation bias of the proportion of incidence due to recent transmission, as inferred from the estimated level of clustering.

3.1 Model Formulation

Following Murray (2002), we consider a population of $N$ people sorted into $n$ clusters. We represent each cluster type by its size ($c$) and assume that a cluster’s size is bounded by a high limit, such that $c = 1, 2, \ldots, C$. With these definitions, a cluster of size one ($c = 1$) includes a DNA-type that is not clustered with any other sample isolates, and is unique.
Similarly, a cluster of size $c$ includes $c$ individuals (person $i=1,2,\ldots,c$) with similar TB isolates. Furthermore, we count the frequency of each cluster type in the population and represent it by $n_c$, where $0 \leq n_c \leq \left\lfloor \frac{N}{c} \right\rfloor$, $c=1,\ldots,C$.

With this definition, each TB isolate can be designated as the $i^{th}$ subject $i=1,\ldots,c$ of cluster, $j=1,\ldots,n_c$, of size $c$. Therefore, the total number of clustered cases can be computed as:

$$\sum_{c=2}^{C} c \times n_c$$

and the proportion of all clustered cases in the population is:

$$\frac{\sum_{c=2}^{C} c \times n_c}{\sum_{c=1}^{C} c \times n_c}$$

We consider an incomplete sampling strategy with a coverage level of $p$, $0 \leq p \leq 1$, where $p$ represents the likelihood of sampling an individual (e.g., $p=1$ represents a complete sampling including everyone in the population, i.e., a census). Therefore, an individual’s sampling events can be represented by independent and identically distributed Bernoulli ($p$) random variables, and the sampling events of individuals from a cluster of size $c$, can be represented using a Binomial($c,p$) random variable. As a result, the expected value of the number of
collected isolates from the whole population in a sampling study with coverage $p$ can be computed as

$$\sum_{c=1}^{C} n_c \times c \times p$$

In such a study, the probability of collecting only one sample from any given cluster of size $c$ can be estimated as

$$c \times p \times (1-p)^{c-1}$$

Therefore, the expected number of unique isolates included in the sample (which will not be counted toward clustering) is

$$\sum_{c=1}^{C} n_c \times c \times p \times (1-p)^{c-1} = n_1p + \sum_{c=2}^{C} n_c cp(1-p)^{c-1}.$$ 

This consists of two terms: the number of originally unique isolates (in the population) successfully included in the sample ($n_1p$); and the number of single isolates taken from original clusters (and therefore mistakenly represented as unique isolates). Finally, the level of clustering in this sample from the population is computed as
which corresponds to the ratio computed by the “$n$” method: the number of clustered cases divided by the sample size. Murray and Alland (2002) develop a very similar model to estimate the ratio of unique strains in the sample (Murray & Nardell 2002). Both models confirm our initial expectations regarding the role of the following factors on the clustering estimation bias:

1. Sampling coverage and study duration determining the incomplete sampling likelihood of $p$.
2. Underlying cluster-size distribution determining $C$ and $n_c$ for each cluster type.

In extending this analysis, we first consider a time-independent setting (such that we’re not concerned with the study duration), and we assume that the sampling coverage determines the value of $p$ for each individual in our population. In a general setting, the underlying distribution of the cluster-size distribution can be stated as a function of existing cluster sizes and the frequency of each cluster ($n_c$) such that the total number of clustered and unique isolates sums to the population size:

$$N = \sum_{c=1}^{C} cn_c.$$ 

Due to uncertainty in the underlying cluster-size distribution, the size of the biggest cluster ($C$) or the frequency of existing clusters ($n_1$ to $n_c$) is not known. Here, we assume a
simplified representation of the underlying cluster-size distribution using a linear interpolation of three major points:

- $n_1 =$ frequency of clusters of size 1
- $n_2 =$ frequency of clusters of size 2
- $n_C =$ frequency of clusters of size $C$

In order to explore the properties of this model, we assume a constant population size of $N$, with a known number of unique clusters ($U$), and a fixed maximum cluster size $C$. Once the total number of unique cases, $U = n_1$ is fixed, the remaining population ($N-U$) can be assigned to clusters of size 2 to $C$, such that

$$N - U = \sum_{c=2}^{C} c n_c,$$
$$n_c \in \left[ 0, \left\lfloor \frac{N-U}{c} \right\rfloor \right].$$

Under these assumptions, however, the shape of the cluster size distribution can still change. Figure 2, top graph, illustrates various examples of possible cluster-size distributions under fixed values of $n_1$ (number of unique cases), $n_2$, and $n_C$. The choice of possible distributions, however, is subject to a hard constraint: cluster sizes ($n_2$, ..., $n_C$) should be chosen such that the total number of clustered cases ($2n_2+3n_3+...+Cn_C$) is equal to $N-U$ where $U = n_1$ represents the number of unique cases (Figure 2, bottom graph).
In this case, we can estimate the distribution of cluster sizes using a linear interpolation of cluster size 2 to \( C \) (fitting a straight line between \( n_2 \) and \( n_C \)). Tuning the value of \( n_2 \) and \( n_C \) will further enable us to approximate various settings of the cluster-size distribution. Figure 3 illustrates examples of various cluster-size distributions using this method. In such cases, the form of the distribution is tied only to the values of \( n_2 \) and \( n_C \), and once these two values are determined, all other intermediate cluster sizes (\( n_3, \ldots, n_{C-1} \)) are computed through linear interpolation of these two points. Based on the choice of \( n_2 \) and \( n_C \), the resulting distribution can range from a right- to a left-skewed distribution.

Figure 2: Various cluster-size distributions under a fixed population of clustered cases and maximum cluster size.
Figure 3: Linear approximation of the underlying cluster size distribution.

Note that in all three cases, the total number of clustered cases $N-U$, and the maximum cluster size $C$ remain constant while the distribution of cluster sizes change. The linear approximation enables computing the frequency of all intermediate cluster sizes ($n_c$, $c=2,..,C$) as a function of $n_2$ and $n_C$:

$$cn_c - 2n_2 = (c-2) \frac{Cn_C - 2n_2}{C-2}$$

$$n_c = \frac{1}{c} \left( (c-2) \frac{Cn_C - 2n_2}{C-2} - 2n_2 \right)$$
Moreover, due to the global constraint on the total number of clustered cases \((N-U)\), \(n_c\) can be stated as a function of \(n_2\):

\[
N - U = \frac{(2n_2 + Cn_c)(C - 1)}{2} \rightarrow n_c = \frac{1}{C} \left( \frac{2(N - U)}{C - 1} - 2n_2 \right)
\]

where \(n_2 \in \left[0, \frac{2(N - U) - C(C - 1)}{2(C - 1)}\right]\) to ensure a minimum value of \(n_c=1\) (i.e., we assume that the biggest cluster size must be always present in the population, or otherwise not stated). This definition enables us to represent various settings of cluster-size distributions by tuning the value of \(n_2\), estimating the corresponding level on \(n_c\), and computing the linear interpolation of intermediate cluster sizes (\(n_3\) to \(n_{C-1}\)), i.e., ranging from a more right-skewed distribution (\(n_2 >> n_c\)), to a uniform (\(n_2=n_c\)), and a left-skewed distribution (\(n_2 << n_c\)) (see Figure 3).

### 3.2 Numerical Analysis

In order to study the sensitivity of clustering-estimation bias to variation in the underlying cluster-size distribution, we now consider a number of numerical examples. The examples are programmed in Matlab (see Section 1 of the Supplementary Material). In each example, we compute the estimation bias of the level of clustering under various settings of the cluster-size distribution and study coverage.
where $TC$ is the true level of clustering in our original population ($TC=1-U/N$). Figure 4-A shows the estimation bias of the level of clustering in a population of 10000 people, with a clustering level of 90%, as a function of $n_2$ and $p$. Part A is a 3D surface plot of bias as a joint function of both $n_2$ and $p$ simultaneously; Part B is a plot of bias as a function of $n_2$ alone for various fixed levels of $p$ on the different lines; and Part C is a plot of bias as a function of $p$ alone for various fixed levels of $n_2$ on the different lines. Clustering-level estimation bias is sensitive to the study coverage and reduces as the proportion of sampled cases increases (Figure 4-C). Moreover, similar samples tend to result in better clustering estimates under lower values of $n_2$, corresponding to a left-skewed cluster-size distribution (Figure 4-B). This suggests that molecular sampling studies tend to work better in the presence of a small number of big clusters (as $n_C$ increases) where there is a lower probability for missing clustered cases and subsequent mis-classification of those for unique strains.
One-way Sensitivity Analysis of Estimation Bias to Maximum Cluster Size:

Figure 5 shows a one-way sensitivity analysis of the estimation bias of clustering-level (estimated level-true proportion of clustered cases) to variation of maximum cluster size (C). The graphs on the top show the estimation bias (y-axis) as a function of $n_2$ (lines representing various fixed levels of $p$ corresponding to the lower graphs), and graphs at the bottom show the estimation bias (y-axis) as a function of $p$ (lines representing various fixed levels of $n_2$ corresponding to the upper graphs). Each pair (top-bottom) compares the behavior under a different value of the maximum cluster size $C=10, 30, 50$ (left to right).
An overall comparison of errors indicates that increasing the maximum cluster size will subsequently reduce the range of estimation bias. This confirms our earlier results for the role of size clusters \( (C=50) \) to reduce the chance of misclassification and improve estimation of the clustering level.

**One-way Sensitivity Analysis of Estimation Bias to the True Clustering Ratio**

Figure 6 shows the one-way sensitivity analysis of the estimation bias to variation of the underlying true clustering level. Each graph illustrates the estimation bias of clustering (y-axis) as a function \( n_2 \) (x-axis). Different graphs compare the behavior under variation of the true clustering ratio \( (TC \text{ changing from 0.95 at top-left to 0.5 at bottom-right}) \). The colored line on each graph shows the level of study incompleteness \( p \) ranging from 10\% to 100\%.
The graphs suggest that similar studies tend to result in better estimates of clustering in populations with lower clustering levels. This pattern confirms our previous findings: increasing the clustering level under a fixed maximum cluster size (fixed $C$) results in creation of more clusters, which in turn increases the chance of mis-classification of clustered cases and therefore increases the estimation bias.

3.3 Discussion

Linear approximation of the cluster-size distribution enables studying the clustering bias in relation to various forms of the cluster-size distribution and study incompleteness. Our results suggest that an increase in the number of DNA clusters will increase the chance of
misclassification (i.e., increased likelihood of a single draw from a small-size cluster, and misclassifying this sample it as a unique strain), and will subsequently increase the underestimation of clustering level. This implies that molecular studies dealing with left-skewed cluster-size distributions have a higher likelihood of misclassifying samples and therefore underestimating the true clustering level.

While our current analysis depicts the general effects of cluster-size distribution on estimation bias of the clustering level, we cannot quantify this bias without knowing the true form of the distribution. The distribution, on the other hand, is an implicit function of epidemic dynamics and host-related characteristics, and is a matter of uncertainty. With regard to this issue, we continue our analysis in the next section, by developing an agent-based simulation model of a TB epidemic that allows us to generate the cluster-size distribution of TB strains, and study the estimation bias of various cluster-analysis methods for the rate of recent transmission in this model.

4 A simulation Approach for Computing the Estimation Bias of the Recent Transmission Rate

We develop an agent-based simulation model of TB transmission, and model the dynamics of strain clustering and survival. Accordingly, our aim here is not to model TB transmission dynamics with precision, but to generate a collection of heterogeneous strain cluster distributions that could be used to demonstrate the effects of sampling. In this section we present the general characteristics of the simulation model and discuss our procedure for calibration, and verification of this model.
4.1 An Agent-Based Simulation Model of a TB Epidemic

We develop an agent based simulation (ABS) model of a TB epidemic in an age-structured population with homogenous mixing. Agents represent people in the simulation model whose behavior and characteristics are modeled at an individual level. The model runs in discrete timesteps representing months, and the outputs are reported at the end of each year. While the definition and duration of the simulation timestep differs across the literature (Mellor et al. 2011; Guzzetta et al. 2011), the choice depends on the goal of the modeling and availability of data for parameterizing the contact pattern (Kasaie et al. 2014). Given the scarcity of data on contact patterns associated with TB and the purpose of this model (to capture the dynamics of TB transmission and strain-diversity over a long period of time), we choose a timestep of one month for our analysis. This timestep is long enough to facilitate crude estimation of monthly social-interaction patterns (described next), yet short enough to reveal the dynamics of TB transmission across a year-long period. The simulation model is coded in C++, and more details are provided in Section 2 of the Supplementary Material.

4.1.1 Population Structure and Contact Network

We consider an age-structured homogeneous populating of 100,000 individuals. The initial population assumes a uniform age distribution (0-90 years old). Individuals are subject to an age-specific annual mortality rate, and die at a maximum age of 90 years. Deceased individual are replaced with a number of newborns at end of each year such that the average population size holds a constant mean over time. The annual number of newborn is therefore generated from a non-stationary Poisson process with a mean value set to the total number of people dying during that year.

The initial model assumes a homogeneous mixing structure in which all members of the population have an equal chance of contact with each other. The contact events are modeled
at the end of each month and for computational efficiency, only the infectious contact events are modeled (i.e., contacts among infectious cases and the rest of the population). We assume that the individual’s number of contacts per timestep (one month) is generated from a Poisson distribution with mean of $\lambda$. The value of $\lambda$ (contact rate) is subsequently tuned to calibrate the incidence rate at the steady state.

4.1.2 TB Natural History

TB natural history is modeled at an individual level using five main TB health states as shown in Figure 7. A person is assumed to be born in full health and susceptible to TB. Upon successful transmission of the disease, the person enters the early latent TB (ELTB) state for a period of five years. During this time, individuals are at high risk of progression to active TB (primary infection); this probability declines with each year in the ELTB state (i.e., the ELTB state actually consists of five sub-states, each lasting one year), according to estimates of adult TB progression following infection (Vynnycky & Fine 1997). After five years post-infection, individuals enter the late latent TB (LLTB) state, during which they have a very low but non-zero probability of progression to active disease in each year (reactivation). Reinfection can occur during early or late latency, and is modeled as a return to year one of the ELTB state; however, individuals with latent TB are assumed to have a degree of immunologic protection from reinfection, such that the probability of infection for an active-to-susceptible contact is higher than the probability of reinfection for an active-to-latent contact.
Patients with active TB (ATB) are infectious and subject to an increased probability of mortality. Infectiousness is modeled as zero at the start of the active period, increasing thereafter as a linear function of time for the first nine months of the disease (i.e., as the bacillary burden grows), and stable thereafter at the maximum level until the individual is diagnosed and treated, or dies. The maximum transmission probability is assumed to be one (full infectiousness), thereby reflecting a combination of intrinsic transmissibility and external factors (e.g., crowding, poverty) that may modify the probability of effective TB transmission. Upon diagnosis and initiation of effective treatment, or alternatively through spontaneous resolution, individuals with active TB enter the non-infectious recovered (REC) state. We calibrate the treatment/recovery rate such that individuals remain infectious for an average of 11 months prior to treatment, as estimated by the WHO (World Health Organization 2013b). Moreover, recovered individuals are subject to a risk of disease relapse in the first two years of the recovery. The annual rate of relapse (per first two years) is tuned to calibrate the proportion of incidence among previously treated individuals. These individuals are also subject to a risk of reinfection, as described above.
4.1.3 Simulating TB Strains

In accordance with the goal of study, we model the diversity and persistence of TB DNA strains among infected cases. In order to create the initial diversity, we assume that all the initially infected individuals carry unique strains of TB (either ELTB, LLTB, ATB or REC). The strains can be transferred from an infectious person to other individuals upon a successful infectious contact, and individuals carry the latest TB strains (i.e., in case of a reinfection, the latest strain replaces the previous existing strains).

Due to the existence of a clustering effect and population birth/death dynamics, however, the initial strain’s diversity will not be sustained for a long time and dies out quickly (Figure 8, darkest red line at the bottom). For example, under no migration rate (closed population with internal births and deaths), no new strain type will enter the population, and after the first 200 years into the model, the number of circulating strains is reduced to a very small number. In contrast, under a higher migration rate of 5% per year, the total number of circulating strains comes to an equilibrium level of about 3000 strains over time (dead strains are replaced with new strains through migration of new individuals to the population). Therefore, we implement a migration process to replenish the number of TB strains such that infected new immigrants carry unique strains of TB. In order to hold a constant population size, we assume an equal rate of migration to and from our population. The migration rate is subsequently tuned to calibrate number of surviving strains through the simulation time.
In order to assist with the computational procedure, we match new immigrants to the population with random members of the population and update their TB strain type upon entering the population (e.g., randomly selected individuals are chosen to leave the population (migrate out), and each one is replaced with another randomly selected member of the population carrying a unique TB strain if infected). Although this procedure may induce an implicit correlation among the disease state levels over time, this is actually beneficial to our analysis by reducing the variability of disease equilibrium-state levels, and population age structure. The migration process takes place at the end of each year, and the actively infected immigrants entering the population are not counted toward disease incidence (since they have acquired the disease prior to entering the population).

4.1.4 Simulation Calibration and Verification

TB parameters are calibrated to the literature and set to mean values used in previously published deterministic models of TB transmission (Table 1). Moreover, we consider a baseline scenario for a setting of medium TB incidence with 200 cases per 100,000/year (representing India) and calibrate the remaining simulation parameters (including the relapse rate and rate of treatment/recovery) in this scenario.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Description/Justification</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Life expectancy</td>
<td>90 y</td>
<td></td>
<td>(World Health Organization 2013a)</td>
</tr>
<tr>
<td>Natural mortality rates</td>
<td>Age-specific mortality rates for India</td>
<td></td>
<td>(World Health Organization 2013a)</td>
</tr>
<tr>
<td>ATB mortality rate</td>
<td>0.12 per year</td>
<td></td>
<td>(Tiemersma et al. 2011)</td>
</tr>
<tr>
<td>Early latency duration</td>
<td>5 years</td>
<td></td>
<td>(Dowdy et al. 2012)</td>
</tr>
<tr>
<td>Cumulative primary infection rate</td>
<td>14.30%</td>
<td></td>
<td>(Vynnycky &amp; Fine 1997)</td>
</tr>
<tr>
<td>Annul primary infection risk (year 1 to 5)</td>
<td>(1, 0.41, 0.13, 0.086, 0.028) *</td>
<td></td>
<td>(Dowdy &amp; Chaisson 2009; Ferebee 1969)</td>
</tr>
<tr>
<td>Reactivation rate</td>
<td>0.001</td>
<td>Calibrated to provide disease duration of 11 months</td>
<td>(World Health Organization 2013b)</td>
</tr>
<tr>
<td>Treatment/Recovery</td>
<td>0.9 per year</td>
<td>Calibrate to provide 15% incidence among previously treated individuals</td>
<td></td>
</tr>
<tr>
<td>Relapse rate (first two years)</td>
<td>0.06 per year</td>
<td></td>
<td>(Kasaie et al. 2014)</td>
</tr>
<tr>
<td>Latent immunity toward re-infection</td>
<td>50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease duration to reach maximum infectiousness</td>
<td>9 months</td>
<td></td>
<td>(World Health Organization 2013b)</td>
</tr>
<tr>
<td>Maximum infectiousness</td>
<td>1</td>
<td>Representing full infectiousness</td>
<td></td>
</tr>
</tbody>
</table>
Using the annual incidence values from the basecase scenario and different settings of TB incidence, we determine a duration of 100 years for the simulation transient period, and start our analysis from year 101 (Figure 9).

![Figure 9: Determining the simulation transient period.](image)

Using the transient period of 100 years, we compute the average output values over 300 years of simulation for the basecase scenario (Table 2). The average output values for the incidence level, susceptibles' risk of infection, ratio of incidence due to primary infection, and proportion of incidence among previously recovered patients agree with the calibration goals.
<table>
<thead>
<tr>
<th>Output (Annul Values)</th>
<th>Average (Standard Deviation)</th>
<th>Description</th>
<th>Calibration Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TB Transmission Dynamics:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total contacts</td>
<td>3430.6 (69.48)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total transmission events</td>
<td>1730.97 (35.68)</td>
<td>Total number of successful TB transmission events</td>
<td></td>
</tr>
<tr>
<td>Susceptible transmissions</td>
<td>1115.15 (22.62)</td>
<td>Number of TB transmissions to susceptible</td>
<td></td>
</tr>
<tr>
<td>Susceptibles' risk of infection</td>
<td>0.02</td>
<td>Annual risk of TB infection for susceptible individuals = (Susceptible transmissions) / (Total number of susceptible)</td>
<td>About 2%</td>
</tr>
<tr>
<td><strong>TB Infection Dynamics:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incidence</td>
<td>200.46 (4)</td>
<td></td>
<td>200 per 100000 in baseline</td>
</tr>
<tr>
<td>Primary infection</td>
<td>142.69 (3.48)</td>
<td>Disease progression from ELTB</td>
<td></td>
</tr>
<tr>
<td>Reactivation relapse</td>
<td>39.52 (1.31)</td>
<td>Disease progression from LLTB</td>
<td></td>
</tr>
<tr>
<td>Ratio of incidence due to primary infection</td>
<td>0.71</td>
<td>Disease activation from REC</td>
<td></td>
</tr>
<tr>
<td><strong>Death/Birth Dynamics:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TB related mortality</td>
<td>23.43 (0.87)</td>
<td>Due to age-specific mortality rate</td>
<td></td>
</tr>
<tr>
<td>Natural mortality</td>
<td>1335.02 (30.58)</td>
<td>When individual's age exceeds 90 years</td>
<td></td>
</tr>
<tr>
<td>Mortality due to old age</td>
<td>153.37 (7.32)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TB States:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SUS</td>
<td>49755.19 (465.58)</td>
<td>Total number of people in each health state</td>
<td></td>
</tr>
<tr>
<td>ELTB</td>
<td>7491.05 (133.61)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLTB</td>
<td>39120.78 (400.77)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATB</td>
<td>194.05 (4.07)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>REC</td>
<td>3439.3 (30.74)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Further Outputs:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population size</td>
<td>100000.37 (6.19)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease duration (months)</td>
<td>11.19 (0.12)</td>
<td>Number of incident cases who has a previous history of treatment/recovery</td>
<td></td>
</tr>
<tr>
<td>Active cases with previous history of recovery</td>
<td>28.13 (1.03)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of incidence among previously recovered patients</td>
<td>0.14</td>
<td></td>
<td>about 15%</td>
</tr>
</tbody>
</table>
4.2 Simulation Experiments Framework

Our earlier results in Section 2 confirmed the previously suggested role of two main elements on the estimation bias of the level of clustering, including sampling incompleteness, and the underlying cluster-size distribution. In this section, we consider various factors describing each element and discuss our experimental framework to study and quantify the effect of each factor on estimation bias of the primary infection ratio.

4.2.1 Experimental Factors

We assume that the sampling of TB isolates is conducted upon TB treatment/recovery, and describe the sampling incompleteness with regard to two factors:

- **Sampling duration** ($d$): the length of the sampling period
- **Sampling coverage** ($p$): the probability of a successful case ascertainment

For the purpose of this analysis, we consider eight levels of variation for study duration, equally distributed over a range of 5 to 40 years of sampling, and assume five levels of variation for the sampling coverage $p$ (see Table 3). In such case, a sampling study with a duration of 5 years and coverage of 20% represents the highest level of sampling incompleteness and tends to miss a great portion of TB patients.

<table>
<thead>
<tr>
<th>Factors</th>
<th>Symbol</th>
<th>Levels</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Incidence</td>
<td>Inc</td>
<td>100</td>
<td>150</td>
<td>200</td>
<td>250</td>
<td>300</td>
<td>350</td>
<td>400</td>
</tr>
<tr>
<td>Ratio of circulating strains</td>
<td>RCS</td>
<td>0.03</td>
<td>0.11</td>
<td>0.19</td>
<td>0.33</td>
<td>0.53</td>
<td>0.63</td>
<td>0.69</td>
</tr>
<tr>
<td>Sampling duration</td>
<td>$d$</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>Sampling coverage</td>
<td>$p$</td>
<td>0.2</td>
<td>0.4</td>
<td>0.6</td>
<td>0.8</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The estimation bias of the cluster analysis method, however, further depends on the underlying cluster-size distribution and epidemic characteristics. Given the uncertainty regarding the true cluster-size distribution, we rely on our simulation-model predictions for the dynamics of TB transmission and survival pattern of TB strains through time. To describe the expected level of clustering in relation to epidemic characteristics, we consider two further factors:

- **TB incidence level**: describing the level of disease transmissibility
- **Ratio of circulating TB strains (RCS)**: describing the diversity of unique strain types.

Since the crude number of circulating TB strains is a function of population size and underlying incidence, we standardize this value with regard to the total infected population size and define RCS as

\[
RCS = \frac{\text{total number of circulating strain types}}{\text{total population of infected people including ELTB, LLTB, ATB and REC}}
\]

RCS levels range between 0 to 1, where a value of 1 corresponds to a population with no DNA clusters such that every infected (or previously infected) person carries a unique strain of TB. The RCS value is, therefore, a function of the migration rate as shown in Figure 10.
The incidence and RCS levels are calibrated by tuning the value of contact-rate and migration-rate parameters in the simulation model. Here, we consider a setting of medium-to-high TB incidence, and assume 8 levels of variation for the underlying incidence ranging from 100 to 450 cases per 100,000/year. Moreover, we consider 8 levels of variation for the RCS level corresponding to the range of the migration rate from 0.5% (representing relatively closed populations) to 10% (representing dynamic population exchange). Table 3 summarizes the information on selected experimental factors and their levels.

### 4.2.2 Design of Simulation Experiments

With regard to the definition of each factor and the simulation model’s logic, we design the simulation experiments on two levels:

1. In the first level, we consider various levels of incidence and RCS resulting in 64 scenarios ($S_1, \ldots, S_{64}$). Each of these scenarios requires a unique initial simulation setting with the corresponding set of input parameters (contact rate $\lambda$ and migration rate $\gamma$). The scenarios are independent, and each can be replicated $R$ times.
2. In the second level, we consider various levels of sampling duration and coverage, resulting in 40 sampling experiments ($e_1, ..., e_{40}$). All experiments should be carried over all replications of 64 scenarios, and the order of experiments can be chosen at random.

In such a setting, we adopt a two-level nested logic for designing the simulation experiment. In level one, we consider the scenarios corresponding to various levels of incidence and RCS (*system factors*). Each simulation scenario is associated with a unique set of simulation input parameters (contact rate and migration rate) and requires a fresh startup of the simulation model (which is computationally rather expensive). Once the simulation model is initialized and past the transient time period, DNA fingerprinting data from all recovered TB patients are recorded into a data repository at the end of each year. In the second level of analysis, we consider various sampling experiments with regard to levels of study duration and coverage (*sampling factors*). The sampling experiments are then carried over the generated DNA repositories and estimate the rate of recent transmission using cluster analysis of TB isolates in each sample. Figure 11 shows the conceptual model of simulation experiments for this analysis. The implementation steps in the analysis are as follow:

1. Generate simulation scenarios based on combinations of system factors: $S_1, ..., S_{64}$.
2. Calibrate the simulation inputs to provide the specified levels of incidence and RCS in each scenario.
3. For each scenario, run the simulation model for $T$ timesteps, and collect the isolates of treated/recovered TB patients at the end of each year. Record the results in the TB isolate repository.
4. Replicate each scenario $R$ times independently.
5. Generate the sampling experiments based on combinations of sampling factors: $e_1, ..., e_{40}$.

235
6. Take each recorded DNA fingerprinting repository from each replication, and carry all sampling experiments over it.

7. Record the experiment’s result for all experiments $(e_1,...,e_{40})$ in each replication $(R)$ of each scenario $(S_1,...,S_{64})$.

Figure 11: Conceptual model of simulation experiments.

For the sampling-experiment logic, we use a repeated-sampling strategy to increase the precision of the sampling experiments’ results. Consider a single replication $r$ of a scenario $s$, where $r=1,...,R$ and $s=1,...,64$. The TB isolate repository corresponding to this replication includes the annual TB isolates of all treated patients in this model: from the beginning of the simulation model to year $T$ (end of the simulation). Moreover, consider a single sampling experiment $e_n$, where $n=1,...,40$. This experiment is associated with a pre-specified level of duration and coverage $(p,d)$ and will be carried over the DNA fingerprint
repository as follows: beginning from year \( t \), we record the DNA isolates of recovered patients with a likelihood of \( p \), and continue the recording procedure for \( d \) years.

Once the sampling is complete, we analyze the collected isolates based on the fingerprint types, and estimate the level of clustering using the ‘\( n \)’ and ‘\( n-1 \)’ methods. In this procedure, the sampling starting time \( t \) can be chosen arbitrarily from year 100 (beginning of the simulation steady state) to year \( T-d \).

An experiment result collected in this fashion, however, is subjected to stochastic variations of TB dynamics throughout the simulation period \( t \) to \( t+d \), which eventually inflates the variability of the experiment’s results and reduces the power of the design. Therefore, we use a random sampling procedure to increase the precision of the reported experiment’s results in the domain of each simulation replication (Figure 12). In this procedure, we repeat the experiment \( e_n \) \( B \) times, each time using a random starting point \( t \), and report the average value of these repeated experiments \( \bar{e}_n \) as the final experiment result. The choice of \( B \) is made such that the relative precision of the final reported results is at worst 5% (see Section 3 of the Supplementary Material).

![Diagram](image)

**Figure 2: Sampling experiments’ logic.**
4.2.3 Calibrating Experimental Settings

Combinations of incidence and RCS levels (system factors) result in 64 simulation scenarios. Each scenario is simulated for \( T \) timesteps, and is independently replicated \( R \) times. In each single replication, we carry out all the sampling experiments (40 experiments). Each experiment is carried out using a random starting point, and is repeated \( B=50 \) times to increase the precision. In such a case, our goal is to choose the simulation time horizon long enough to generate a sufficiently big DNA repository such that the experiment results in the same replication do not correlate with each other. This issue arises due to the computational expense of simulation initialization: it’s cheaper to run a single replication for 1000 years than to replicate a model of 100 years 10 times.

In order to perform a factorial analysis of the results, however, it’s required that the amount of variation for experiment results inside a single replication (\( \text{Var}(e^r) \), where \( e^r = (e^r_1, e^r_2, ..., e^r_{40}) \)) is not statistically different across several replications (\( \text{Var}(e^i) = \text{Var}(e^j) \) where \( i \) and \( j \) are independent replications of a given scenario).

In this analysis, we consider \( R=10 \) replications and assume a time horizon of \( T=1000 \) years, which provides 900 years of annual TB DNA fingerprint data for our sampling experiments (i.e., skipping the initial transient period in each model). Details on the statistical analysis of this assumption are provided in Section 3 of the Supplementary Material.

5 Results

Analysis of system factors: Each simulation scenario corresponds to a unique setting of system factors (incidence and RCS), which are calibrated by tuning the value of the contact-rate and migration-rate parameters in the simulation model. Figure 13 shows the relationship
between the average incidence values and variation of contact rate and migration rate across the simulation scenarios. Each value is computed as the mean annual incidence throughout the simulation steady state (year 100 to 1000), and is averaged across the replications of each scenario.

![Incidence pattern across simulation scenarios](image)

**Figure 13:** Incidence pattern across simulation scenarios: A. Incidence (vertical axis) as a function of both migration rate and contact rate simultaneously (horizontal axes), B. Projection of incidence on contact-rate-axis (different levels of migration overlap each other and are not shown), C. Projection of incidence on migration-rate-axis (lines representing different levels of contact rate).

Incidence values show a rather linear relationship with the monthly contact rate of disease: a higher contact rate generates more infectious contacts, and boosts the incidence. Incidence shows negligible sensitivity to variation of the migration rate.
Similarly, the sensitivity of RCS levels to variation of incidence and migration rate is shown in Figure 14. While the RCS level is primarily controlled by the migration rate, it is also sensitive to the underlying incidence level such that higher incidence levels correspond to lower RCS: when the rate of disease transmission increases, the clustering effect is empowered, and the total ratio of surviving strains declines.

![Figure 14: RCS pattern across simulation scenarios: Left graph shows RCS values as a function of migration rate (colored lines represent various levels of incidence). Right graph shows RCS values as a function of incidence level (colored lines represent various levels of migration rate).](image)

Finally, we study the true rate of primary infection in each scenario. The true value of primary infection is computed as the proportion of disease incidences due to primary progressions for ELTB. Figure 15 shows the pattern of changes for the true primary infection ratio under variations of the incidence levels.
The ratio of disease due to primary infection increases (almost linearly) with the underlying incidence level. This relationship is especially useful for interpreting the rate of primary infection from cluster-analysis results. These results suggest that under similar settings of population structure and disease-related attributes (e.g., rates of primary infection, reactivation, etc), higher disease incidence is associated with a higher rate of primary infection. Therefore, fixed sampling studies results for the clustering estimates can have different interpretations for the level of primary infection based on the underlying incidence level.

**Sampling Experiments:** Combinations of sampling factors (sampling duration and coverage) generate 40 experiments. All experiments are carried over all replications of each scenario. Moreover, each experiment is repeated $B=50$ times over each TB isolate repository (see Section 3 of the Supplementary Material). i.e., choosing a random starting time, drawing a random sample of population (based on coverage $p$), analyzing the sample, and repeating the procedure 50 times. Finally, the average value of collected results (across the 50 repetitions) is reported as the final observation for that sampling experiment. In analysis of TB isolate samples, we consider several outputs as shown in Table 4.
Table 4: Cluster analysis outputs.

<table>
<thead>
<tr>
<th>Outputs</th>
<th>Description</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y0</td>
<td>Sample size</td>
<td>$N_s$</td>
</tr>
<tr>
<td>Y1</td>
<td>Number of unique strains</td>
<td>$U$</td>
</tr>
<tr>
<td>Y2</td>
<td>Number of clustered strains</td>
<td>$N_s - U$</td>
</tr>
<tr>
<td>Y3</td>
<td>Number of clusters (with 2 or more members)</td>
<td>$n$</td>
</tr>
<tr>
<td>Y4</td>
<td>Number of secondary case</td>
<td>$N_s - U - n$</td>
</tr>
<tr>
<td>Y5</td>
<td>Primary Infection estimate by &quot;n&quot; method</td>
<td>$(N_s - U)/N_s$</td>
</tr>
<tr>
<td>Y6</td>
<td>Primary Infection estimate by &quot;n-1&quot; method</td>
<td>$(N_s - U - n)/N_s$</td>
</tr>
</tbody>
</table>

5.1.1 Effect of Incidence and RCS on Clustering Estimate

Figure 16 shows the sensitivity of the number of unique strains, clustered strains, and the total number of clusters to variation of incidence and RCS. The results are presented for only a single experiment setting with the most representative sampling study ($e_{40}$: duration of 40 years and coverage of 100%).
As the diversity of circulating strains increases, the frequency of clusters (both unique strain types (clusters of size 1), and clustered strain types) increases (Figure 16, A2 & A3). However, the pattern of change is proportional to the underlying incidence level: the magnitude of changes increases with the incidence level (Figure 16, B2 & B3).

The total number of clustered cases, on the other hand, shows a close dependence on level of incidence: higher transmission empowers the clustering effect and increases the total number of clustered cases regardless of the underlying migration rate. The distribution of these clusters however, depends on the level of migration: a higher migration rate results in a larger number of small clusters.

Finally, Figure 17 compares the results of sampling studies for levels of clustering under different analysis methods (‘n’ vs. ‘n-1’). The results are presented for only a single sampling experiment with a duration of 40 years and 100% coverage. As expected, the

Figure 16: Sensitivity of clustering measures to RCS and incidence levels.
clustering estimate by the ‘\(n-1\)’ method is generally lower than with the ‘\(n\)’ method, and the clustering estimate declines as the diversity of circulating strains increase. The clustering estimate in both methods is insensitive to variation in incidence, which in turn can be a problem for interpreting the rate of primary infection from the clustering level: under different incidence levels, equal clustering levels can correspond to different rates of primary infection (see Figure 17).

![Figure 17: Clustering estimate by ‘\(n\)’ vs. ‘\(n-1\)’ method.](image)

**5.1.2 Estimation Bias for Ratio of Incidence due to Recent Transmission**

In this section, we focus on the results of the ‘\(n-1\)’ clustering method for estimating the level of primary infection. The estimation bias is computed as the difference of the estimated clustering level minus the true rate primary infection.

Figure 18 compares the changes of estimation bias across various experiments with different levels of coverage and duration. In an ideal study setting with long duration and complete coverage \((d=40, p=1)\), the estimation bias is very sensitive to the diversity of
circulating strains (Figure 18 left): low strain diversity corresponds to a small number of big-size clusters, which lowers the chance of misclassification toward unique strains. In such a case, the clustering method tends to overestimate the primary infection ratio. As the strain diversity increase, there is a higher chance of mis-classifying smaller clusters toward unique strains, and the estimated clustering level (and inferred primary infection rate) decreases.

In contrast, decreasing the study duration and coverage will lower the average clustering level estimate, and underestimates the rate of primary infection (Figure 18, right). A similar behavior is observed with regard to the changes of strain diversity: higher diversity results in greater underestimation of primary infection rate.

In such case, we further study the pattern of bias under various levels of strain diversity. Figure 19 shows eight estimation-bias surfaces under different levels of RCS (eight levels, increasing from top-left to bottom-right). Each graph illustrates the estimation bias as a function of both study coverage and duration jointly, and all scenarios use a fixed incidence level of 200 per 100,000/year.
The overall behavior of the response in each graph implies more severe underestimation for lower levels of study coverage and duration (e.g., an incomplete study has a higher likelihood of missing clustered cases, and underestimating the level of recent transmission).

Moreover, a general comparison of all eight graphs implies that increasing the strain diversity results in more severe underestimation of the recent transmission rate (i.e., the average response level is decreasing uniformly as RCS increases). Similarly, we consider the behavior of estimation bias in relation to various levels of incidence (Figure 20), under a fixed RCS level of 30%.

**Figure 19: Estimation bias of primary infection in relation to RCS.**
A general comparison of all graphs suggests that higher incidence levels are associated with relatively worse estimation bias (overestimation). This can be explained with regard to the role of incidence in determining the cluster-size distribution: higher transmission levels empower the clustering effect (formation of a small number of large-size clusters), which in increases the average estimates of clustering level.

5.1.3 Reponses-Surface Analysis

As mentioned before, an ultimate goal of this study is to develop a relatively simple and easy-to-use (in the field) decision-support tool for adjusting the estimated rate of recent transmission in future molecular studies of TB. Our simulation results enable us to measure the estimation bias (from its true value) in a variety of settings. In this section, we aim to unify our findings and develop a simple algebraic predictive model of approximate estimation bias as a function of our decision factors.
A simulation analysis is often exercised when cost, time, or other constraints prohibit experimentation with the real system. In such case, experimentation with a computer simulation model allows for understanding the system’s performance, and assists in behavior prediction. In some circumstances, however, experimenting with a simulation model may still not be an efficient (or even feasible) option due to long simulation run times, and further approximations are required. In such cases, simpler approximations can be made: deterministic models of simulation models, called *metamodels* (Kleijnen 2008). Metamodels, or *response-surface models*, are simple deterministic approximating functions of the simulation output, and are based on first- or second-order polynomials regression in the form of

\[ y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \varepsilon, \quad \varepsilon \sim i.i.d. N(0, \sigma^2), \]

or

\[ y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} x_i x_j + \varepsilon, \quad \varepsilon \sim i.i.d. N(0, \sigma^2), \]

where the \( \beta_i \)s are the estimated regression coefficients, for simulation output \( y \) as a function of the decision factors \( x_i \)s (for notational simplicity, we refrain from putting the “hats” on the \( \beta_i \)s as is sometimes done to signal that they are sample-based estimates rather than population-based parameters, but our \( \beta_i \)s are sample-based estimates).

In this section, we perform a response-surface analysis of our main output, the estimation bias of the ‘\( n-1 \)’ cluster-analysis method for the rate of recent transmission, with regard to the four decision factors of our analysis: incidence, RCS, sampling duration (\( d \)),
and sampling coverage ($\rho$). Our objective is to develop a simple deterministic predictive regression model, with satisfactory fit to the actual simulated data, which is easy to disseminate and apply in future molecular studies of TB. While the simulation model provides a more flexible and realistic representation of the system, direct distribution and application of this model by non-modelers may not be practical. Therefore, we try to approximate our results through a deterministic regression model, which is easy to distribute and apply in analysis of future DNA fingerprinting data. The major issues in developing such model, however, include the choice of a functional form for the regression model, fitting the regression model to the simulation response using the experimental data, and assessment of the adequacy of the fit (Barton & Mechesheimer 2006).

In the following, we present our analysis procedure and results for developing a regression metamodell of the simulation model, as well as discuss the advantages and limitation of this approach. The analysis is performed in SAS and details are provided in Section 4 of the Supplementary Material.

**Initial Analysis: A First-Order Model of Main Effects**

We start our analysis by fitting a first-order (linear) model of the main four effects:

$$Y_{ijklm} = \beta_0 + \beta_i A_i + \beta_j B_j + \beta_k d_k + \beta_l p_l + e_{ijklm},$$

where

- $i = 1, \ldots, 8$
- $j = 1, \ldots, 8$
- $k = 1, \ldots, 8$
- $l = 1, \ldots, 5$
- $m = 1, \ldots, 10$
where A and B stand for incidence and RCS (each with 8 levels), $m$ is the number of simulation replications, the $\beta_i$s are the regression coefficients, $Y_{ijklm}$ is the a single simulation-generated observation for the estimation bias of the recent transmission rate, and $\epsilon_{ijklm}$ is the error term. Table 5 shows the results of analysis of variance for this model.

### Table 5: Analysis of variance for the first-order model of main effects.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>570.34635</td>
<td>142.58659</td>
<td>81570.0</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>25595</td>
<td>44.74074</td>
<td>0.00175</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack of Fit</td>
<td>2555</td>
<td>44.44236</td>
<td>0.01739</td>
<td>1343.14</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pure Error</td>
<td>23040</td>
<td>0.29838</td>
<td>0.00001295</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>25599</td>
<td>615.08709</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results of Analysis of Variance table and the lack of fit test imply a significant fit of the selected model to data. Furthermore, individual t-tests of regression coefficients are all significant (see Section 4 of the Supplementary Material). Also, we inspected the fit diagnostics graphs to confirm meeting the required regression assumptions for a normal distribution of independent residuals with homogenous variance in each subgroup (Figure 21).
The plot of residuals against predicted values shows an apparent curvature, which can be an indicator of higher-order effects. This violates the assumption of identical distribution of residuals and requires refining the regression model.

Selecting the Best Second-Order Global Regression Model

Our initial analysis’s results suggest the presence of higher effects in the model. This is not surprising with regard to the large sample size (about 25,600 observations in total) and multiple levels of each factor that result in a powerful ANOVA test (with a very high degree of freedom) enable to detect very small changes in variance.
In refining the initial model, we include additional polynomial terms for higher powers of main effects up to the 4\textsuperscript{th} order (\(x_i^2\): quadratic, \(x_i^3\): cubic and \(x_i^4\): quartic terms), as well as the two-way (\(x_ix_j, i,j=1,...,n, i \neq j\)) and three-way (\(x_ix_jx_k, i,j,k=1,...,n, i \neq j \neq k\)) interactions.

This results in inclusion of 38 factors (4 one-way, 10 two-way, 20 three-way, and 4 four-way effects) in the regression model. In order to find the best polynomial regression among possible combinations of these factors, we set our main selection criteria to be minimizing the adjusted R\textsuperscript{2} (\(R_{adj}^2\)) value and compare candidate models with regard to several criteria of goodness of fit (see Section 4 of the Supplementary Material for more information on the selection criteria). Table 6 shows the results of analysis for the best three selected models.

<table>
<thead>
<tr>
<th>Model Index</th>
<th>Number in Model</th>
<th>Adjusted R-Square</th>
<th>Adjusted R\textsuperscript{2}</th>
<th>C(p)</th>
<th>AIC</th>
<th>BIC</th>
<th>Estimate of Predicted MSE</th>
<th>Root MSE</th>
<th>SBC</th>
<th>S(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35</td>
<td>0.9975</td>
<td>0.9975</td>
<td>33.8</td>
<td>-248703</td>
<td>-</td>
<td>248701</td>
<td>0.0001</td>
<td>0.000077</td>
<td>248409</td>
</tr>
<tr>
<td>Variables in Model:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>0.9975</td>
<td>0.9975</td>
<td>35.7</td>
<td>-248701</td>
<td>-</td>
<td>248699</td>
<td>0.0001</td>
<td>0.000077</td>
<td>248399</td>
</tr>
<tr>
<td>Variables in Model:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>0.9975</td>
<td>0.9975</td>
<td>35.7</td>
<td>-248701</td>
<td>-</td>
<td>248699</td>
<td>0.0001</td>
<td>0.000077</td>
<td>248399</td>
</tr>
<tr>
<td>Variables in Model:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Criteria of fit for best model: results are only presented for the first three selected models.
The first suggested model to maximize the value $R^2_{adj}$ is composed of 35 terms. Note that the amount of $R^2_{adj}$ variation among the first three models is negligible and the only apparent difference is due to very small changes of other fit measures (C(p) and AIC). Figure 22 shows the fit criteria for selecting this model based on the main six measures of fit.

![Figure 22: Fit criteria for selecting the best model.](image)

Moreover, we double-check this selection through a *forward selection method*, using SAS *Proc Reg* (Beal 2007). This method begins from the simplest regression model with only the intercept term. For each of the independent variables, the ANOVA $F$-statistic is calculated to determine each variable’s contribution to the model, and only the variable with the smallest $p$-value (below a specified cutoff probability) is kept in the model. In order to avoid inflation of experiment-wise error in the multiple comparisons, we conservatively set the
cutoff probability to small error of 0.001 per comparison. Figure 23 compares the fit criteria of the candidate models in this analysis.

![Fit Criteria for Y](image)

**Figure 23:** Fit criteria for best regression model in forward selection.

The final selected model using the forward selection corresponds to the previously suggested polynomial regression with the same 35 terms, providing additional credibility to this regression metamodel. Table 7 shows the parameter estimation and $t$-test of significance for each coefficient in this model.
**Table 7: Parameter estimates for the selected polynomial model coefficients.**

| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|-------|----|-------------------|----------------|---------|------|---|
| Intercept | Intercept | 1 | -0.19998 | 0.00233 | -85.73 | <.0001 |
| A | A | 1 | 0.00017733 | 5.1E-06 | 34.62 | <.0001 |
| B | B | 1 | -1.79106 | 0.00563 | -317.9 | <.0001 |
| d | d | 1 | 0.03314 | 0.00019 | 175.29 | <.0001 |
| p | p | 1 | 1.26261 | 0.01617 | 78.1 | <.0001 |
| A3 | 1 | -1.02E-10 | 1.37E-11 | -7.45 | <.0001 |
| B2 | 1 | 1.9893 | 0.02376 | 83.73 | <.0001 |
| B3 | 1 | -1.15002 | 0.04719 | -24.37 | <.0001 |
| B4 | 1 | 0.18597 | 0.00019 | 175.29 | <.0001 |
| d2 | 1 | -0.00148 | 1.4E-05 | -102.17 | <.0001 |
| d3 | 1 | 0.0003414 | 4.67E-07 | 73.15 | <.0001 |
| d4 | 1 | -3.08E-07 | 5.16E-09 | -59.61 | <.0001 |
| p2 | 1 | -1.99128 | 0.02376 | 83.73 | <.0001 |
| p3 | 1 | 1.63792 | 0.06583 | 28.81 | <.0001 |
| p4 | 1 | -0.52336 | 0.02364 | -22.14 | <.0001 |
| AB | 1 | 0.00060483 | 1.3E-05 | 48 | <.0001 |
| Ad | 1 | -0.0006942 | 1.96E-07 | -48.04 | <.0001 |
| Bp | 1 | 0.93933 | 0.005 | 187.99 | <.0001 |
| Bd | 1 | 0.01063 | 0.0012 | 90.14 | <.0001 |
| Ap | 1 | -0.00041514 | 1.2E-05 | -35.32 | <.0001 |
| pd | 1 | -0.01305 | 0.0011 | -117.34 | <.0001 |
| Apd | 1 | 7.01E-07 | 1.31E-07 | 5.36 | <.0001 |
| p2B | 1 | -0.33241 | 0.00282 | -118.01 | <.0001 |
| p2d | 1 | 0.00327 | 0.02364 | 51.63 | <.0001 |
| p2A | 1 | 0.00016778 | 6.3E-06 | 26.51 | <.0001 |
| B2A | 1 | -0.00099515 | 9.5E-06 | -104.46 | <.0001 |
| d2A | 1 | 8.15E-08 | 3.70E-09 | 22.05 | <.0001 |
| d2p | 1 | 0.00012017 | 1.5E-06 | 80.24 | <.0001 |
| ABd | 1 | 0.00006688 | 1.44E-07 | 47.91 | <.0001 |
| B2d | 1 | -0.01211 | 9.5E-05 | -127.07 | <.0001 |
| B2p | 1 | -0.37241 | 0.00386 | -96.52 | <.0001 |
| d2B | 1 | -0.00006689 | 1.7E-06 | -40.64 | <.0001 |
| Bdp | 1 | 0.00176 | 5.8E-05 | 30.25 | <.0001 |
| ABp | 1 | 0.00015819 | 5.8E-06 | 27.19 | <.0001 |
| A2p | 1 | -5.95E-08 | 1.46E-08 | -4.08 | <.0001 |
| A2B | 1 | 2.23E-07 | 1.62E-08 | 13.76 | <.0001 |
The $p$-value of the $t$-test is significant for all estimated parameters and confirms their significance to stay in the regression model. Next, we inspect the fit diagnostics graphs to confirm meeting the required regression assumptions, as shown in Figure 24.

![Fit diagnostics for second-order polynomial model.](image)

The graph of residuals versus the predicted values shows no apparent curvature pattern that could be an indicator of yet higher-order higher effects. The residuals follow a normal distribution and the $R^2_{adj}$ value of 0.9975 indicates a satisfactory fit to the data. Further
analysis of residual plot versus regressors, however, suggests a pattern of change for variance of residuals in each subgroups (Figure 25). The plots of residuals against sampling coverage ($p$) and duration ($d$) shows an apparent pattern of variance heterogeneity across different levels of these factors. i.e., higher variance for lower levels of $p$ ($p=0.2$) and $d$ ($d=5$ years). Such a pattern violates the assumption of *homoscedasticity* (homogeneity of variances) per the requirement of the method of ordinary least squares (and of the ANOVA test) used to fit the regression model. Further analysis of a BoxCox transformation of the data does not change this conclusion.

![Residual by Regressors for Y](image)

**Figure 25:** Plots of residuals by regressors.

In such a case, one solution is to ignore the pattern of heteroscedasticity (in favor of a very large sample size like the one we have), and continue with the fitted regression. Moderate deviations from the assumption of equal variances do not seriously affect the
results in the analysis of variance, and the usual ANOVA $F$-test is relatively robust when the groups are all about the same size.

An alternative solution is to try method of \textit{generalized least square} (GLS), instead of the ordinary analysis of variance, to adjust the matrix of variances and fit a new metamodel under the assumption of heteroscedasticity. Due to computational difficulties in developing such a model, however, the initial tradeoff between the convenience and performance of the metamodel does not hold (e.g., a GLS model is computationally expensive and yet does not guarantee better performance than the ANOVA model or simulation model). Therefore, we drop this alternative in favor of other options.

A final alternative is to break the data into pieces within each of which we have homoscedastic variance, and fit local metamodels to each subgroup. A further inspection of simulated data with regard to various levels of each main factor (incidence, RCS, study coverage, and study duration) implies a significant pattern for variance of data across levels of $p$ and $B$ (representing RCS), as shown in Figure 26.
Therefore, an alternative is to break the data into smaller subgroups with regard to levels of these factors, repeat the test to ensure homoscedasticity of data within each subgroup, and fit a local metamodel to the data within each group. In such a case, the outcome presents a collection of locally fitted metamodels that should better approximate the actual simulation model. One problem with this procedure, however, is uncertainty regarding the definition of each sub-group, which can require a long computational procedure to define them. In the presence of a large number of sub-groups, the initial tradeoff between the convenience and performance of such metamodels may not hold. Finally, the actual amount of improvement in the predictive power of local metamodels in comparison to the global metamodel is not guaranteed. This is potentially an area of future research in extending the current analysis.

Figure 26: Box-plots of response versus main effects.
Discussion

An ultimate goal of this study is to develop a decision-support tool for adjusting the estimated rate of recent transmission in future molecular studies of TB. Due to practical issues regarding the distribution and direct application of simulation model by non-modelers, we approximate our results through a simple deterministic regression model, which is easy to distribute and apply in future molecular studies of TB (it could even be implemented in a just spreadsheet).

Despite heterogeneity of residuals, the suggested polynomial regression does a good job of approximating the actual simulation model’s behavior, and can be used as a predictive model for estimation bias of recent transmission rates: researchers just plug in the regression variables for levels of TB incidence, strain diversity (RCS), study duration and coverage, and compute the regression response for estimation bias. Adding this value to the current estimate of clustering level (using the ‘n-1’ method) will adjust the final estimate of the proportion of incidence due to recent transmission.

An alternative approach, and a preferred one, is direct application of the simulation model to compute the estimation bias. The ABS model provides a flexible and yet powerful platform to simulate the dynamics of TB transmission, and to generate the underlying cluster size distribution overtime. Furthermore, the simulation allows for modeling different DNA sampling studies to analyze the clustering level and estimating the ratio of incidence due to recent transmission. A direct application of this model by non-modelers, however, will require further expansion of the current model to incorporate a user-friendly interface, an optimization engine to calibrate the simulation input (contact rate and migration rate) to specified levels of incidence and RCS, a database for collecting many replications of the model to ensure a certain level of precision, and an output-delivery system for summarizing.
6 Conclusion

In this research, we study the estimation bias of the ratio of incidence due to recent transmission in epidemiological studies of DNA fingerprinting data. We propose a mathematical model of clustering-level estimation, which provides insight into the general behavior of clustering estimators in the presence of different cluster-size distributions and sampling-study coverage. Moreover, we propose a simulation-based approach to study the role of various factors on the estimation bias of the recent transmission rate. We include four main factors in relation to the epidemic structure and sampling characteristics, including the disease incidence level, the ratio of total circulating strains in the population (RCS), sampling duration, and sampling coverage. Using an ABS model of TB as a virtual laboratory, we implement a series of statistically controlled experiments with regard to combinations of factor levels and compute the estimation bias in each setting.

Our findings agree with the results of previous studies for the behavior of clustering-level estimators with regard to variation of each factor (Glynn, Vyonycky, et al. 1999; Murray, Megan 2002). Clustering-level estimation depends on the underlying cluster-size distribution, and the underestimation bias increases as the frequency of small clusters increases (representing a left-skewed cluster-size distribution). This effect is observed in the presence of high RCS values (corresponding to a large number of circulating strains) and lower incidence levels (slower cluster formation). Moreover, the estimation of the clustering level is a function of study coverage and duration, and the bias increases with the level of incompleteness (smaller coverage and shorter duration). In such cases, the sampling study tends to miss a larger frequency of TB cases in a similar chain of transmission, either due to
inefficient ascertainment of active cases, or short duration of the study to allow for disease activation.

Moreover, interpretation of the estimated clustering level for the ratio of incidence due to recent infection varies with the underlying incidence of disease. Our results suggest that, under similar settings of population structure and disease-related parameters (e.g., rates of primary infection, reactivation, etc.), higher disease incidence is associated with a higher rate of primary infection (and recent transmission). Therefore, similar clustering-level estimates corresponding to different incidence settings will have different levels of correspondence to the true rate of recent transmission. A direct translation of clustering level to the recent transmission rate (as is usually done in the cluster-analysis method) is therefore another source of bias.

The scope of the current analysis is of course limited to the maximum and minimum level of each factor representing a setting of medium-to-high TB incidence (100 to 450 cases per 100,000/year), a ratio of circulating strains ranging from 2% to 70%, and a study duration from 5 to 40 years. Further analysis is required to expand the scope of results to values of these ranges. Due to uncertainty regarding the actual level of strain diversity (RCS), the researcher may choose to use a range of RSC values to compute an expected range of bias. In general, however, it is recommended to use a higher estimate of the RCS level that is more realistic (see below). Finally, the simulation analysis relies on the assumption that the epidemic system is at a steady-state equilibrium state. Further analyses are required for settings of low TB incidence (e.g., in the US and Canada), and non-equilibrium states with decreasing or increasing incidence levels.

The current analysis relies on a modeling assumption for using migration as a means to replenish the number of circulating TB strains through time. An alternative approach used in a previous study (Vynnycky et al. 2001) is modeling the mutation process of TB strains in
which each strain will have a probability of mutating to a new (unique) strain through time. One problem with this approach is the uncertainty regarding the mutation rate of TB strains. Moreover, recent epidemiological advances in DNA fingerprinting allow for identification of close generations of mutated strains, and further decreases the chance of misclassification of mutated strains. An alternative solution is modeling a source of external transmission that would result in infection with unique TB stains. This modeling approach is more complex than is the current migration-based logic, and requires a more complicated analysis of surviving strains through time. We consider this as a future research direction for expansion of the current study. Preferably, we expect to observe similar results under both modeling approaches, which would further confirm the fidelity of the current migration-based modeling approach.

An ultimate objective of this study is to develop a decision-support tool to be used in future epidemiological studies of TB. We present different alternatives for this purpose. One option is to use a deterministic metamodel of the estimation bias as a function of four decision factors, as described in Section 5.1.3. Such a model is easy to disseminate and apply, but in turn, inherits a higher level of modeling error and does not guarantee satisfactory performance under different scenarios. Another option is direct application of ABS model, which provides a flexible and more powerful platform to measure the estimation bias under a variety of settings. This will in turn, require further expansion of the simulation model to incorporate a user-friendly interface, as well as a computational infrastructure for handling the input calibration and output estimation, and is an area of future development.
7 References


CDC, 2011. A CDC Framework for Preventing Infectious Diseases: Sustaining the Essentials and Innovating for the Future,


Chao, D., 2013. Calibrating an influenza epidemic model.


266


Holm, J., 1969. *Development from tuberculous infection to tuberculous disease*. TSRU Progress Report;KNVC; The Hague, The Netherlands,


Vynnycky, E. et al., 2001. The effect of age and study duration on the relationship between`clustering'of DNA fingerprint patterns and the proportion of tuberculosis
disease attributable to recent transmission. *Epidemiology and infection*, 126(1), pp.43–62.


Supplementary Material

1. Analytical Model of Clustering Estimation

MATLAB code for computing the estimation bias of clustering level using a linear approximation of the cluster-size distribution (The MathWorks 2014):

```matlab
function outputBias=plotBias()
N =10000;    %sample size
U  % percentage of unique samples
C  % maximum cluster size
p  % sample coverage
n2 % frequency of cluster size 2
nC % frequency of cluster size C
p = 0:.1:1;           % a row-vector of p
U=.1;
C= 50;
Step=10;
Max=(2*(N-U*N)/(C-1) -C)/2;
n2 = 1 :Step:  Max  ;          % a row-vector of b
[X,Y] = meshgrid(p,n2); % create matrices for grids of them
Z=arrayfun(@(computeClusteringBias,X,Y);
hl=surf(X,Y,Z);
myStr=strcat('C=',num2str(C));
colormap hsv
xlabel 'p';
ylabel 'n2';
zlabel 'Estimation Bias';
title (myStr);
end

function ClusteringBiase = computeClusteringBias(p,n2)
nC =((2*(N-N*U)/(C-1)) -2*n2)/C;
if (nC<0)
    ClusteringBiase=NaN;
else
    %compute expected number of unique strains in the sample
    num=U*N*p;
    for j=2:1:C
        num=num+ (((j-2)*(C*nC-2*n2)/(C-2)) +2*n2)* p*(1-p)^(j-1);
    end
    %compute total expected number of strains in the sample:
    dom=U*N*p;
    for j=2:1:C
        dom=dom+ (((j-2)*(C*nC-2*n2)/(C-2)) +2*n2)* p;
    end
    %compute the clustering estimate bias
    ClusteringBiase =(1- num/dom)-(1-U));
    end
end
```
2. ABS Model Details

The simulation model is programmed in C++, and includes five main classes. Agents represent people in the transmission model, and are defined in class “Person.cpp”. The initial population forms a single community and is created at the beginning of the simulation using class “Community.cpp”. The simulation model is directed from the Main class “driver.cpp”; all parameters are globally defined in “Params.cpp”, and additional methods are provided in “methods.cpp”.

An individual’s initial health state is determined with regard to a specified endemic TB equilibrium which, in turn, is calibrated to the equilibrium level of a corresponding deterministic transmission model (Kasaie et al. 2013). The initial age structure is assumed to follow a uniform distribution (between ages 0 to 90).

The model runs in discrete time steps representing months, and the outputs are recorded at the end of each year. The natural population dynamics and disease-transmission events are modeled by the following methods, programmed at the community level:

- **Contact method**: called at monthly intervals
  - Generating the random contacts of infectious individuals through a random draw from a Poisson distribution with mean \( \lambda \).
  - Applying the probability of transmission upon each contact based on an individual’s infectiousness and immunity levels.
  - Modeling successful transmission events: proceeding to the ELTB state.

- **Progress method**: called at monthly intervals
  - Managing the disease progression among previously infected individuals.
  - Checking the current time against the time to leave the current state, and progressing to the next state if true.
Death /birth method: called at yearly intervals

- Modeling the natural mortality events across the population due to age-specific mortality rates or reaching the maximum age
- Modeling the birth events for a random number of newborns (tuned to keep the average population size constant): the number is determined from a random draw from a Poisson distribution with mean set to the difference between the current population size and the initial population size.

Migration method: called at yearly intervals

- Selecting a random number of population members to migrate out
- Replacing those leaving the population, with random members of the population whose TB strain is updated to a new strain

Here, the disease progression is modeled at the individual level, via a countdown of the remaining time to progression to the next state. These designated event times to leave each state are determined with regard to the TB transmission rates, and are generated at the arrival to that state. Once a person enters a health state (e.g., LLTB), the associated exit routes from that state are identified (e.g., slow progression to ATB), and the random remaining time to execute those events are computed with regard to the chance of events (e.g., the remaining time to slow progression is set to a random draw from a geometric distribution with mean 0.001/12 ). At the end of each timestep, the progress method checks the current model time against the remaining time to exit each state, and executes events that are due.

The simulation random number-generator is chosen from GLS (the GNU Scientific Library). GSL is a numerical library for C and C++ programmers, and it is free software under the GNU General Public License. The random-number generator algorithm (gsl_rng_taus2) is based on a maximally equi-distributed combined Tausworthe generator.
by L’Ecuyer with a period of $2^{88}$ (about $10^{26}$) (L’Ecuyer 1996), with an improved seeding procedure compared to the original edition (L’Ecuyer 1996).

3. Calibrating the Experimental Settings

A full factorial experiment setting requires i.i.d observations. In our analysis, each observation corresponds to a sampling experiment on a given scenario: $8 \times 8 \times 8 \times 5 = 2560$, and in theory each observation should be collected from an independent simulation run. Due to the stochastic nature of simulation, we further need to collect several replications of each observation ($2560 \times R$), which increases the required number of independent replications substantially. Due to the computational cost associated with initializing each independent run (and skipping the transient period), we choose a two-level experimental logic as discussed in section and run $8 \times 8 \times R = 64 \times R$ independent replications of the simulation model. Each model simulates for $T$ years, and generates corresponding DNA fingerprinting data with a length of $T - 100$ years.

In such a case, our goal is to choose the simulation time horizon long enough to generate a big DNA repository such that the experimental results collected from the same replication do not correlate with each other. This requires that the amount of variation for the experimental results inside a single replication ($\text{Var}(\epsilon')$, where $\epsilon' = (\epsilon'_1, \epsilon'_2, \ldots, \epsilon'_{40})$, $r=1, \ldots, R$) is not statistically different across several replications ($\text{Var}(\epsilon') = \text{Var}(\epsilon')$ where $i$ and $j$ are independent replications of a given scenario).

In our initial analysis, we consider $R=10$ replications and assume a time horizon of $T=1000$ years. Graphical comparison of estimation bias ($Y$) across replications, do not indicate an apparent difference in means (Figure 1).
Moreover, Analysis of variance of the experiments’ results (\(Y=\) estimation bias by the \(n-1\) method) in each scenario do not indicate a significant effect for replications (Table 1).

![Distribution of Y](distribution.png)

**Figure 1: Overall comparison of estimation bias (Y) across 10 replications.**

In addition, we perform an overall test for analysis of variance with regard to all four factors of interest (\(A=\)incidence, \(B=\)RCS, \(p=\)coverage, \(d=\)duration) and replications (Table 2 and 3). The results confirm the statistically insignificant role of replications.

### Table 1: Analysis of variance for replication effect in an arbitrarily scenario corresponding to incidence =100 and RCS = 2%.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>9</td>
<td>0.001525</td>
<td>0.000169</td>
<td>0.1</td>
<td>0.9995</td>
</tr>
<tr>
<td>Error</td>
<td>390</td>
<td>0.629426</td>
<td>0.001614</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>399</td>
<td>0.630951</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>REP</td>
<td>9</td>
<td>0.001525</td>
<td>0.000169</td>
<td>0.1</td>
<td>0.9995</td>
</tr>
</tbody>
</table>
Table 2: Analysis of variance for effect of main factors and replication.

<table>
<thead>
<tr>
<th>Class</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>REP</td>
<td>10</td>
<td>0 1 2 3 4 5 6 7 8 9</td>
</tr>
<tr>
<td>p</td>
<td>5</td>
<td>0.2 0.4 0.6 0.8 1</td>
</tr>
<tr>
<td>d</td>
<td>8</td>
<td>5 10 15 20 25 30 35 40</td>
</tr>
<tr>
<td>A</td>
<td>8</td>
<td>100 150 200 250 300 350 400 450</td>
</tr>
<tr>
<td>B</td>
<td>8</td>
<td>0.02 0.1 0.18 0.32 0.50 0.61 0.68 0.73</td>
</tr>
</tbody>
</table>

Number of Observations Read 25600

Number of Observations Used 25600

The ANOVA Procedure
Dependent Variable: Y Y

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>34</td>
<td>589.4665</td>
<td>17.33725</td>
<td>17394.3</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>25565</td>
<td>25.48111</td>
<td>0.000997</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>25599</td>
<td>614.9476</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Significance test for regression coefficients.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Anova SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>4</td>
<td>106.6285952</td>
<td>26.6571488</td>
<td>26744.9</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>d</td>
<td>7</td>
<td>36.9762796</td>
<td>5.2823257</td>
<td>5299.72</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>A</td>
<td>7</td>
<td>0.2749850</td>
<td>0.0392836</td>
<td>39.41</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>445.5860118</td>
<td>63.6551445</td>
<td>63864.7</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>REP</td>
<td>9</td>
<td>0.0006320</td>
<td>0.0000702</td>
<td>0.07</td>
<td>0.9999</td>
</tr>
</tbody>
</table>
Choosing the Number of Experiment’s Re-Samples (B)

We use a repeated sampling strategy to increase the precision of the sampling experiments’ results. In this procedure, we repeat each experiment, \( e_{n}^{r,s} \) \( n=1,\ldots,40; \ r=1,\ldots,10; \ s=1,\ldots,64 \) (corresponding to the \( n \)th experiments carried out over the \( r \)th replication of scenario \( s \)), \( B \) times, each time using a random starting point \( t \), and report the average value of these repeated experiments \( e_{n}^{r,s} \) as the final experiment result:

\[
\overline{e_{n}^{r,s}} = \frac{B}{\sum_{b=1}^{B} e_{n,b}^{r,s}}
\]

The choice of \( B \) is made such that the relative precision of the final reported results is at worst 5%:

\[
\varepsilon = \frac{h_{0.95}(e_{n}^{r,s})}{e_{n}^{r,s}} \leq 0.05
\]

where \( h_{0.95}(e_{n}^{r,s}) \) is the 95% confidence interval around mean of \( e_{n}^{r,s} \). We compute the precision of the results of all the reported outputs of the cluster-analysis method (see Table 4), and the results confirm that the average relative precision is no worse than 5%.
4. Response-Surface Analysis

We use SAS Proc REG to perform the linear regression analysis (SAS Institute Inc. 2013). The general linear statistical model for the factorial design data is in the form of

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} \quad i = 1, \ldots, n; j = 1, \ldots, n_i, \varepsilon_{ij} \sim NID(0, \sigma^2) \]

where, \( n \) is the number of independent variables, and \( n_i \) is the number of observations in each group. Adding higher-order terms (quadratic or interaction terms) to this model results in

\[ y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 + \beta_3 x_i x_j + \varepsilon_{ij} \quad i = 1, \ldots, n; j = 1, \ldots, n_i, \varepsilon_{ij} \sim NID(0, \sigma^2) \]

which is still a linear regression model; i.e., although the expressions on the right hand side are quadratic in the independent variables \( x_i \), it is still linear in the parameters \( \beta_i \). In both cases, the \( \varepsilon_{ij} \) s are the error terms assumed to be independent and identically distributed normal random variables with mean 0 and constant standard deviation \( \sigma \). Given a random sample from the population, then, we estimate the population parameters and obtain the estimated linear regression model. The regression parameters can be estimated through the

<table>
<thead>
<tr>
<th>Average precision across all runs</th>
<th>( \varepsilon_1 )</th>
<th>( \varepsilon_2 )</th>
<th>( \varepsilon_3 )</th>
<th>( \varepsilon_4 )</th>
<th>( \varepsilon_5 )</th>
<th>( \varepsilon_6 )</th>
<th>( \varepsilon_7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.023151</td>
<td>0.033953</td>
<td>0.025556</td>
<td>0.031031</td>
<td>0.027232</td>
<td>0.004413</td>
<td>0.006638</td>
</tr>
</tbody>
</table>

Table 4: Average output precision across all runs.
method of *ordinary least square* (OLS), which aims to minimize the sum of squared residuals, $SSE$, by solving a system of *normal equations* resulting from this model:

$$Obj : Min(SSE) = Min \sum_{i=1}^{T} \sum_{j=1}^{n} \varepsilon_{ij}^2, \quad \varepsilon_{ij} = y_{ij} - \hat{\theta}_0 - \hat{\theta}_0 x_{ij}.$$ 

Once a regression is constructed, we check the goodness of fit of the model and statistical significance of the estimated parameters (through the ANOVA F test).

**Initial Analysis: A Linear Model of Main Effects**

We begin our analysis by fitting a simple linear model of the main four effects (incidence, RCS, $d$, and $p$) to the simulation output (estimation bias of the recent transmission ratio), as described in Section 5.1.3. Tables 5-7 below show SAS output for analysis of this model.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>570.34635</td>
<td>142.58659</td>
<td>81570.0</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>25595</td>
<td>44.74074</td>
<td>0.00175</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack of Fit</td>
<td>2555</td>
<td>44.44236</td>
<td>0.01739</td>
<td>1343.14</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pure Error</td>
<td>23040</td>
<td>0.29838</td>
<td>0.00001295</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Corrected Total</strong></td>
<td>25599</td>
<td>615.08709</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6: Summary measures for the linear model of main effects.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root MSE</td>
<td>0.04181</td>
</tr>
<tr>
<td>R-Square</td>
<td>0.9273</td>
</tr>
<tr>
<td>Dependent Mean</td>
<td>0.02006</td>
</tr>
<tr>
<td>Adj R-Sq</td>
<td>0.9272</td>
</tr>
<tr>
<td>Coeff Var</td>
<td>208.36989</td>
</tr>
</tbody>
</table>

Table 7: Parameter estimation for the linear model of main effects.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td>1</td>
<td>0.03413</td>
<td>0.00109</td>
<td>31.22</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>1</td>
<td>-0.00002840</td>
<td>0.00000228</td>
<td>-12.45</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td>1</td>
<td>-0.51153</td>
<td>0.00102</td>
<td>-503.87</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>p</td>
<td>1</td>
<td>0.21765</td>
<td>0.00092385</td>
<td>235.59</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>d</td>
<td>1</td>
<td>0.00295</td>
<td>0.00002281</td>
<td>129.42</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The p-values for analysis of variance, and lack of fit tests are very small, and indicate that the model is able to explain the majority of variances in the data and provides a good fit. Moreover, the statistical tests of the parameter’s estimates show that all four parameters are significant and should remain in the model.

Moreover, the results indicate a relative high value for R-squared, the coefficient of determination, or $R^2$. $R^2$ measures the proportion of the variability of the dependent variable that is explained by the regression model:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$
where SSR is the sum squares of variance explained by the regression model, and SST is the total variance of the data. The adj R-square (\( R_{adj}^2 \)) is a modification of \( R^2 \) that adjusts for the number of factors in a model relative to the number of data points, and measures the proportion of the variability of the dependent variable that is explained by the variation of the independent variables after accounting for the intercept and the number of independent variables:

\[
R_{adj}^2 = 1 - \frac{(N - b) \cdot SSE}{(N - p) \cdot SST},
\]

where \( N \) is the number of observations, \( p \) is the number of independent variables in the model, and \( b \) is a the binomial variable \( b \) (\( b = 1 \) if the intercept is included in the model, \( b = 0 \) otherwise). Unlike \( R^2 \), \( R_{adj}^2 \) increases when a new term is included only if the new term improves the \( R^2 \) more than would be expected in the absence of any explanatory value’s being added.

**Selecting the Best Model**

Maximizing \( R_{adj}^2 \) can be used as the selection criterion to find an ideal combination of several factors in a regression model providing the best fit and without excess/unnecessary terms. In our study, however, the large sample size (\( N = 25600 \)) can decrease the ability of \( R_{adj}^2 \) to distinguish between similar models (close values of \( p \)). Therefore, while we set the
main selection criterion based on maximizing $R_{adj}^2$, we compare and rank similar models with regard to other criteria of fit including:

- **Mallows’ Cook (Cp)** statistic is a measure of total squared error defined as
  
  \[
  Cp = \frac{SSE_p}{\sigma^2} - (N - 2p)
  \]
  
  where $SSE_p$ is the sum-of-squared errors for a model with $p$ parameters, and $\sigma^2$ is the pure variance in the full model (MSE).

- **Akaike’s information criterion (AIC)** defined as
  
  \[
  AIC = n \ln \left( \frac{SSE}{N} \right) + 2p .
  \]
  
  The first term is a measure of the model lack of fit, while the second term ($2p$) is a penalty term for additional parameters in the model. The model with the smallest AIC is deemed the best.

- **The Bayesian Information Criteria (BIC)** is defined as
  
  \[
  BIC = n \ln \left( \frac{SSE}{N} \right) + \frac{2(p + 2)N\sigma^2}{SSE} - \frac{2N^2\sigma^4}{SSE^2}
  \]
  
  where $\sigma^2$ is the pure variance in the full model. The penalty term of BIC is more complex than in the AIC.

We refer the reader to (Beal 2007) for more information on other selection criteria and their applications.

**Analyzing the Model Fit**

Tables 8-10 show the ANOVA output and parameter estimation in fitting a second-order regression model with 35 terms, as described in Section 5.1.3.
Table 8: ANOVA results for second-order regression model.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>35</td>
<td>613.54586</td>
<td>17.52988</td>
<td>290764</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>25564</td>
<td>1.54123</td>
<td>0.00006029</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack of Fit</td>
<td>2524</td>
<td>1.24285</td>
<td>0.00049241</td>
<td>38.02</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pure Error</td>
<td>23040</td>
<td>0.29838</td>
<td>0.00001295</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>25599</td>
<td>615.08709</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Fit criteria for second-order regression model.

| Root MSE         | 0.00776 | R-Square | 0.9975 |
| Dependent Mean   | 0.02006 | Adj R-Sq | 0.9975 |
| Coeff Var        | 38.69726 |          |        |
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|-------|----|-------------------|----------------|---------|--------|
| Intercept | Intercept | 1 | -0.19998 | 0.00233 | -85.73 | <.0001 |
| A | A | 1 | 0.00017733 | 0.00000512 | 34.62 | <.0001 |
| B | B | 1 | -1.79106 | 0.00563 | -317.90 | <.0001 |
| d | d | 1 | 0.03314 | 0.00018907 | 175.29 | <.0001 |
| p | p | 1 | 1.26261 | 0.01617 | 78.10 | <.0001 |
| A3 | 1 | -1.0234E-10 | 1.37284E-11 | -7.45 | <.0001 |
| B2 | 1 | 1.98930 | 0.02376 | 83.73 | <.0001 |
| B3 | 1 | -1.15002 | 0.04719 | -24.37 | <.0001 |
| B4 | 1 | 0.18597 | 0.03083 | 6.03 | <.0001 |
| d2 | 1 | -0.00148 | 0.00001444 | -102.17 | <.0001 |
| d3 | 1 | 0.00003414 | 4.667306E-7 | 73.15 | <.0001 |
| d4 | 1 | -3.07665E-7 | 5.161677E-9 | -59.61 | <.0001 |
| p2 | 1 | -1.99128 | 0.04739 | -42.02 | <.0001 |
| p3 | 1 | 1.63792 | 0.05685 | 28.81 | <.0001 |
| p4 | 1 | -0.52336 | 0.02364 | -22.14 | <.0001 |
| AB | 1 | 0.00060483 | 0.00001260 | 48.00 | <.0001 |
| Ad | 1 | -0.0000942 | 1.961129E-7 | -48.04 | <.0001 |
| Bp | 1 | 0.93933 | 0.00500 | 187.99 | <.0001 |
| Bd | 1 | 0.01063 | 0.00011795 | 90.14 | <.0001 |
| Ap | 1 | -0.00041514 | 0.00001175 | -35.32 | <.0001 |
| pd | 1 | -0.01305 | 0.00011118 | -117.34 | <.0001 |
| Apd | 1 | 7.006617E-7 | 1.307237E-7 | 5.36 | <.0001 |
| p2B | 1 | -0.33241 | 0.00282 | -118.01 | <.0001 |
| p2d | 1 | 0.00327 | 0.00006328 | 51.63 | <.0001 |
| p2A | 1 | 0.00016778 | 0.00000633 | 26.51 | <.0001 |
| B2A | 1 | -0.00099515 | 0.00000953 | -104.46 | <.0001 |
| d2A | 1 | 8.151555E-8 | 3.69743E-9 | 22.05 | <.0001 |
| d2p | 1 | 0.00012017 | 0.00000150 | 80.24 | <.0001 |
| ABd | 1 | 0.00000688 | 1.436466E-7 | 47.91 | <.0001 |
| B2d | 1 | -0.01211 | 0.0009526 | -127.07 | <.0001 |
| B2p | 1 | -0.37241 | 0.00386 | -96.52 | <.0001 |
| Variable | Label | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|-------|----|--------------------|---------------|---------|-------|
| d2B      |       | 1  | -0.00006689        | 0.00000165    | -40.64  | <.0001 |
| Bdp      |       | 1  | 0.00176            | 0.00005818    | 30.25   | <.0001 |
| ABp      |       | 1  | 0.00015819         | 0.00000582    | 27.19   | <.0001 |
| A2p      |       | 1  | -5.95004E-8        | 1.457344E-8   | -4.08   | <.0001 |
| A2B      |       | 1  | 2.231879E-7        | 1.622423E-8   | 13.76   | <.0001 |

The ANOVA test indicates a significant fit model to the data, and individual t-tests for the regression coefficients indicate significance of all terms. Moreover, we inspect plots of residuals versus regression terms, which show an apparent pattern of heteroscedasticity with regard to levels of $p$, $d$, and their higher-order effects (Figure 2 below).
Testing the Homoscedasticity Assumption

In order to test the assumption of homoscedasticity (equality of variances) per the requirements of the ordinary least-squares method, we use Levene’s test for homogeneity of
variances, provided in SAS Proc GLM (). We test the null hypothesis (equality of variances) with regard to each main effect \((A, B, p, d)\) separately, and summarize the results in Table 11 below.

### Table 11: Levene's Test for Homogeneity of Y Variance in Proc GLM

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>7</td>
<td>0.1771</td>
<td>0.0253</td>
<td>30.9</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>25592</td>
<td>20.95</td>
<td>0.00082</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>0.249</td>
<td>0.0356</td>
<td>280.63</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>25592</td>
<td>3.2442</td>
<td>0.00013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>4</td>
<td>1.6547</td>
<td>0.4137</td>
<td>1122.87</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>25595</td>
<td>9.4292</td>
<td>0.00037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>7</td>
<td>0.1908</td>
<td>0.0273</td>
<td>40.95</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>25592</td>
<td>17.0333</td>
<td>0.00067</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The p-values of all four tests imply a significant difference in the levels of response variance and reject the assumption of homoscedasticity. Moderate deviations from the assumption of equal variances, however, do not seriously affect the results in the analysis of variance, and the usual ANOVA \(F\)-test is relatively robust when the groups are all about the same size. In such a case, one solution is to ignore the pattern of heteroscedasticity (in favor of a very large sample size like the one we have), and continue with the fitted regression. Alternatively, we can break the data into smaller subgroups with regard to levels of these factors, repeat the test to ensure homoscedasticity of data within each subgroup, and fit a local metamodel to the data within each group. This analysis is a potential area of future research in extending the current study.

### 5. References


Closing Perspectives

“If we had an extra $50 billion to put to good use, which problem would we solve first?”

This is the question put to the participants of the Copenhagen Consensus in 2004 (Lomborg 2006). The expert panel included eight top economists who were asked to provide in a list of priorities of the world’s most challenging problems. Ranked first on the expert panel’s list was controlling HIV/AIDS. At a cost of $27 billion, around 28 million cases of the illness could be prevented by 2010 with a benefit-cost ratio of about 40 times the investment. Hunger, and trade liberalization were number two and three on the list, followed by control and treatment of malaria at the fourth place.

It is not surprising that topping the list of priorities are problems affecting billions of people worldwide, but what may come as a surprise is how basic these problems are. New technology, economic growth, and development have improved living conditions for many people across the world. Nevertheless, there remain the pressing – and basic – needs for adequate food supplies, clean water, and the expectation of good health.

Despite the enormous improvements in global health in the second half of the twentieth century, life expectancy has not increased equally around the world and health disparities are more apparent than ever before. The economic impact of communicable disease can range from microeconomic effects at the household-level to macroeconomic effects on the country-level, and is measured in terms of direct and indirect costs (e.g., the effect of disease on people’s health and quality of life, dollars spent on treatment, lost productivity, etc.). While the microeconomics of ill-health affects almost every country around the world, the burden is especially high in low- and middle-income countries, where limited available funds are insufficient to provide even the most basic health services, and the healthcare costs continue to rise with new technologies and procedures. More than 98% of childhood
deaths occur in developing countries, where communicable diseases represent seven out of the top ten causes of child death, and account for more than 5 million deaths annually. In such a context, one can realize the significant burden of communicable diseases on a global scale, and explain their high ranking on the Copenhagen Consensus list.

Realizing the scope of the problem, however, is only the first step in targeting solutions. In a limited-resource setting, providing the best health care requires effective methods for management and improvement of health-care systems. Healthcare policymaking involves a complex chain of planning, prioritization, and decision making, which in turn demands input and consideration from various branches of medicine, science, engineering, and business. This leads to a wide array of cross-sectional studies and multidisciplinary research on various aspects of health system’s design, implementation, and improvements. In the midst of this research stream, operations and business analytics (OBA) have significant contributions to a variety of applications.

OBA techniques, tools, and theories have long been applied to a wide range of issues and problems in industry, and their important contribution to healthcare is increasingly being recognized. Diverse applications of OBA methods in healthcare range from operations-management implications (e.g., hospital capacity planning, healthcare facility locations), to public policy and economic analysis (e.g., healthcare resource allocation, policy evaluation, and risk management), and clinical applications (Brandeau et al. 2004).

In such a context, my research lies at the intersection of OBA and a specific branch of public-health policymaking for controlling the epidemics of infectious disease. Using a simulation-oriented methodology for modeling and analysis of disease, my research takes advantage of a combination of OBA tools and techniques to address a variety of problems concerning disease dynamics, evaluating control/prevention interventions, allocation of epidemic-control resources, etc., as presented through this dissertation.

293
In choosing the topics of research, I followed three criteria for defining the goals, and refining the scope of each study. The first criterion concerns the contribution of the research, favoring topics with a limited history of research, or specific instances previously not addressed in the literature. For example, the timing of TB infections across various contact networks remains a matter of uncertainty, and despite its important implications for policymaking, very limited research is available on the topic (see Chapter 2). In such a setting, we propose a simulation approach to capture the dynamics of transmission in a generalizable population with representative social structures, and develop one of the first estimates of infection-time distributions across various networks. The findings were subsequently used to address policymaking questions on effectiveness of various case-finding strategies for controlling TB.

The second criterion concerns the effectiveness of the suggested methodology in addressing the problem. While agent-based simulation provides a flexible platform for studying a wide range of social, biological, or industrial systems, one should also note the tradeoff regarding the computational expense of designing, calibrating, and applying ABS models. In such a case, I target problems with especially complex structures (in terms of epidemiological definitions, social structures, stochastic processes, etc.) that do not allow for efficient application of traditional and more convenient mathematical-modeling techniques. The problem of epidemic-control resource allocation (RA) is a good example of such a topic. While traditional approaches to the epidemic RA problem suffer from restrictive assumptions to facilitate exact analytical/mathematical solutions, a simulation-based technique enable us to address a class of freeform epidemic RA problems with minimal modeling restrictions or over-simplifying assumptions.

Finally, the last criterion concerns my personal interest in topics at a middle-ground of theory and practice. Simulation-oriented studies, by nature, are in the direction of applied
research, and the literature contains many instances of pure-applied models (e.g., data-driven simulation models of previous outbreaks). Despite the great importance and practical impact, such studies do not contribute to the theory, and the scope of results is limited to the domain of application. In such a setting, my work targets problems with limited (or often unavailable) theoretical solutions, and proposes simulation-oriented solutions that can extend the scope of theoretical knowledge. Using generalizable designs validated against literature and theory enables us to develop a realistic simulation framework with a high level of power and flexibility. This subsequently allows us to study various aspects of a problem, to investigate a wide range of possible scenarios, and eventually to deduce micro- and macro-level patterns and behaviors not previously studied in the literature. Further extensive sensitivity analysis of parameters and assumptions will extend the scope of results and support application of findings in a variety of settings. For example, in a recent study on estimation bias of the proportion of TB incidence due to recent transmission in epidemiological studies of TB, we explore the range of potential error under a variety of experimental settings with regard to levels of various factors, and ultimately develop a relatively simple predictive model that can serve as a decision-support tool for adjusting the estimate of recent transmission rates in future applications.

Finally, in terms of methodology, the evolution of simulation methods across various studies presented in this dissertation reflects my programming skills and learning curve during the last years of my PhD program. Beginning from a more conceptual framework, my initial research borrowed instances of ABS models from the literature (e.g., Flute simulation model of influenza (Chao et al. 2010)). Later works took advantage of commercial simulation packages (e.g., AnyLogic (XJ Technologies 2013)) to develop models of TB epidemics, which were finally replaced by independently developed applications coded in C++ for parallel simulation and large-scale computation.
In summary, the purpose of the Copenhagen Consensus was to bridge the tower of research and the general public, so that future energy and knowledge are put to good use. In such a context, my dissertation concerns critical global-health issues in understanding, controlling, and policy-making concerning infectious-disease epidemics, and offers a multidisciplinary approach to such problems using advanced computer-simulation techniques and analytical tools. The ABS approach is a novel technique that is increasing in popularity across the literature and in several fields. This brings to bear the power and effectiveness of such models in various applications, and their promising contributions for control and policymaking of infectious diseases.
References


2 Citations are presented for Introduction and Closing Perspective Sections


Lomborg, B. (2006). How to spend $50 billion to make the world a better place.


