Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Jonathan Paul Smith Date

Inference of Inter-Individual Heterogeneity in Tuberculosis Transmission

By

Jonathan Paul Smith
Doctor of Philosophy

Epidemiology

Neel Gandhi, MD
Advisor

Michael Kramer, PhD
Advisor

David Benkeser, PhD
Committee Member

Benjamin Lopman, PhD
Committee Member

Theodore Cohen, MD, DPH, MPH
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Inference of Inter-Individual Heterogeneity in Tuberculosis Transmission

By

Jonathan Paul Smith
M.P.H., Yale University, 2011

Advisor: Neel Gandhi, MD
Advisor: Michael Kramer, PhD

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Epidemiology
2020

Abstract

Inference of Inter-Individual Heterogeneity in Tuberculosis Transmission
By Jonathan Paul Smith

Objectives: Increasing evidence suggests that tuberculosis (TB) transmission is largely characterized by "superspreading," an extreme manifestation of inter-individual heterogeneity wherein a disproportionately small number of individuals contributes to the majority of secondary cases. Superspreading greatly undermines public health interventions and has a profound impact on disease emergence and outbreak trajectory. However, traditional methods used to quantify the propensity for superspreading in a population cannot be applied to TB since high resolution data describing individual-level TB transmission are rarely observed. Fortunately, recent advancements in genotyping have afforded surveillance systems the ability to more accurately identify TB transmission clusters, defined simply as the total number of cases in a given transmission chain. The overall goal of this dissertation was to develop, evaluate, and apply a novel method to quantify the propensity for superspreading in TB using transmission cluster distributions, without the need for more resource-intensive individual data.

Methods: In the first study we utilized branching process theory and a negative binomial offspring distribution to develop a novel method that infers inter-individual heterogeneity using only cluster level data. We then validated the inference procedure under real-world scenarios that lead to imperfect surveillance. In Study 2, we applied this method to TB surveillance data systematically abstracted from the literature and investigated the impact such heterogeneity had on transmission dynamics. In Study 3 we obtained empirical TB surveillance data from the United States Centers for Disease Control and Prevention (CDC) and estimated the propensity for superspreading in four of the most populous states in the US.

Results: Study 1 demonstrated the inference procedure was robust and inferred the same degree of inter-individual heterogeneity as more resource intensive individual-level data. In Study 2, the inferred parameters indicated a similarly high propensity for superspreading across various global contexts. Study 3 demonstrated a similarly high propensity of superspreading the US, and that a small minority ($\sim 10\%$) of cases were responsible for all secondary transmission.

Conclusions: A high degree of inter-individual heterogeneity is a defining feature of TB epidemiology, and accounting for this heterogeneity in epidemic modeling will result in an improved understanding of TB transmission dynamics and subsequent public health efforts.

Inference of Inter-Individual Heterogeneity in Tuberculosis Transmission

By

Jonathan Paul Smith
M.P.H., Yale University, 2011

Advisor: Neel Gandhi, MD
Advisor: Michael Kramer, PhD

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Epidemiology
2020

Acknowledgments

For their dedication, support, and unwavering patience, I would like to thank my dissertation committee for their commitment to my education and assistance with this body of work. I would also like to recognize Dr. Andrew Hill and Dr. Benjamin Silk at the Centers for Disease Control and Prevention for their commitment to ensuring this work both came to fruition and is integrated in ongoing epidemiological applications. I appreciate the true commitment to education throughout the entire faculty and staff in the Department of Epidemiology at Emory University's Rollins School of Public Health. I am deeply grateful for my parents, Nancy and John Smith, as well as my wife, Dr. Tara Ariel Streich-Tilles, all of whom have provided unconditional love and support throughout my entire educational career.

Finally, I dedicate this work to my daughter, Charlotte Anne Smith. Looking into her eyes will always make it impossible to give up.

Contents

1 Introduction and Background **1**

1.1 Dissertation Approach and Summary 1

1.1.1 Study 1: Specific Aims and Summary 3

1.1.2 Study 2: Specific Aims and Summary 4

1.1.3 Study 3: Specific Aims and Summary 4

1.2 Tuberculosis Epidemiology and Transmission Dynamics 6

1.2.1 Global Tuberculosis Epidemiology 6

1.2.2 Inter-Individual Heterogeneity and Superspreading in Tuberculosis Transmission 9

1.2.3 Sources of Inter-Individual Heterogeneity in TB Transmission 11

1.3 Current Approaches to Addressing Heterogeneity in TB Transmission 17

1.4 Chapter 1 References 21

2 Quantifying Inter-Individual Heterogeneity in Tuberculosis Transmission **32**

2.1 Background 32

2.2 Branching Process Analysis 35

2.2.1 Branching Process Overview 35

2.2.2 Incorporating Inter-Individual Heterogeneity in Branching Process Models 36

2.3 Parameter Inference from the Distribution of Final Transmission Cluster Sizes 40
2.3.1 Relating the Individual Offspring Distribution and the Final Cluster Distribution 41
2.4 Chapter 2 References 45

3 Evaluating a Method to Infer Inter-Individual Heterogeneity in TB transmission Using Cluster Level Data 47
3.1 Abstract 47
3.2 Background 49
3.3 Methods 50
3.3.1 Statistical Methods 50
3.3.2 Maximum Likelihood Estimation of Transmission Parameters 53
3.3.3 Simulated Data 53
3.3.4 Complications in TB Surveillance 54
3.3.5 United States National TB Surveillance System Data 55
3.4 Results 56
3.4.1 Initial Validation the Inference Procedure 56
3.4.2 Bias Arising Due to Complications in Surveillance 57
3.4.3 Validation of Inference Procedure Under Real World Scenarios 58
3.4.4 Analysis of United States TB Surveillance Data 60
3.5 Discussion 60
3.6 Supplemental Materials 74
3.7 Chapter 3 References 82

4 Estimates for the Propensity of Superspreading in Tuberculosis Transmission from Global Surveillance Systems 87
4.1 Abstract 87

4.2 Introduction . . . 89
4.3 Methods . . . 90
4.3.1 Search Strategy . . . 90
4.3.2 Inclusion and Exclusion Criteria . . . 91
4.3.3 Parameter Inference Using Cluster-level Data . . . 91
4.3.4 Cluster Size Probability Calculations . . . 92
4.4 Results . . . 93
4.4.1 Characteristics of Included Datasets . . . 93
4.4.2 Transmission Parameter Estimates . . . 93
4.4.3 Cluster Size Probabilities . . . 94
4.5 Discussion . . . 95
4.6 Supplemental Materials . . . 104
4.7 Chapter 4 References . . . 105

5 Estimating individual heterogeneity in tuberculosis transmission in the United States 108

5.1 Abstract . . . 108
5.2 Introduction . . . 109
5.3 Methods . . . 111
5.3.1 Data Source . . . 111
5.3.2 Inference Procedure and Model . . . 111
5.3.3 Transmission Cluster Definitions . . . 113
5.3.4 Burden of Secondary Transmission . . . 113
5.4 Results . . . 114
5.5 Discussion . . . 115
5.6 Supplemental Materials . . . 125
5.7 Chapter 5 References . . . 130

6 Summary and Conclusions **132**

6.1 Overview 132

6.2 Public Health and Epidemiological Implications 134

6.3 Limitations 135

6.4 Remaining Gaps in Knowledge and Future Directions 137

Appendix A Key Formulas **141**

A.1 Probability Density Function 141

A.2 Likelihood Equation 141

Appendix B Relevant R Code for Inference Procedure **142**

B.1 Branching Process Function 142

B.2 Imperfect Simulation Function 143

B.3 Likelihood Function 151

B.4 Parameter Estimation Function 153

List of Figures

1.1 Global cases of tuberculosis, 2000-2018 6
1.2 TB morbidity and mortality are related to country income 8
1.3 Pathways of disease occurrence after inhalation of *Mtb* 12

2.1 Visualization of individual- vs cluster-level data of TB transmission in a surveillance system 34
2.2 The dispersion parameter k and its impact on transmission dynamics 39

3.1 Common complications arising in TB transmission surveillance 65
3.2 Estimates of k by case ascertainment probabilities 67
3.3 Impact of censoring on estimates of k 68
3.4 Bias arising due to overlapping clusters 69
3.5 Performance of inference procedure under imperfect surveillance . . . 71
S3.1 Inference of k under various R values for individual- and cluster-level data 74
S3.2 Inference of k under various R values under perfect surveillance, under varying values of N 75
S3.3 Inference of R under varying case ascertainment probabilities 76
S3.4 Visualizing the bias of missing cases 77
S3.5 Surface Estimates of R and k in the United States 78

S3.6 Partial rank correlation coefficients (PRCCs) comparing imperfect surveillance parameters with inference of transmission parameters 79
S3.7 Coverage probabilities of empirical estimates of R and k using US CDC TB surveillance data 81

4.1 Joint estimates of the reproductive number R and dispersion parameter k for included studies 100
4.2 Joint estimates of the reproductive number R and dispersion parameter k for included studies 102
S4.1 Relationship between R, k, and the probability of a cluster size of at least 15 cases 104

5.1 Transmission cluster size distributions in the U.S. states of California, Florida, New York, and Texas 120
5.2 Joint estimates of R and k by state 122
5.3 Proportion of infected individuals responsible for 80% of the total secondary transmissions (p_{80}) across the current consensus range of R values for TB in the United States 124
S5.1 Joint estimates of R and k by state, with clusters defined at the state level 127
S5.2 Proportion of infected individuals responsible for 80% of the total secondary transmissions (p_{80}) across the current consensus range of R values for TB in the United States (State-level definitions) 129

List of Tables

1.1 Brief summary of epidemiologic models used in assessing infectious disease transmission dynamics 18

3.1 Individual vs cluster-level inference of k under simulated R and k values 66

3.2 Parameter values for simulated scenarios representing high, moderate, and low resource settings. 70

3.3 Transmission cluster sizes in the United States by timeframe and geography 72

3.4 Estimates of R and k for TB transmission in the United States by timeframe and geographic definition of clusters 73

S3.1 Coverage probabilities of the inference procedure under various k values 80

4.1 Characteristics of Included Studies 98

4.2 Cluster size distribution of included surveillance datasets 99

4.3 Cluster-based maximum likelihood estimates of $\hat{R}$ and $\hat{k}$ 101

4.4 Absolute and relative probability of observing largest cluster in observed data 103

5.1 Reported TB cases and clusters in the United States, 2014-2016 . . . 119

5.2 Maximum likelihood estimates of R and k, by state 121

5.3 Expected percent of transmission attributed to a given proportion of the cases, p_t 123

S5.1 Cluster size distributions of TB in the U.S., by cluster definitions . . 125

S5.2 Maximum likelihood estimates of R and k, by state, with transmission clusters defined at the state level 126

S5.3 Expected percent of transmission attributed to a given proportion of the cases, p_t (Cluster definitions defined at state level) 128

Chapter 1

Introduction and Background

1.1 Dissertation Approach and Summary

This body of work is comprised of three studies which develop, evaluate, and employ a novel method to quantify the extent of inter-individual heterogeneity in tuberculosis (TB) transmission, defined as differences in the number of secondary cases arising between infectious individuals within a population. With more than 10 million new cases and 1.5 million deaths in 2018, TB is a major contributor of global morbidity and mortality.[1] Incident cases of TB arise in a population from either sporadic reactivation of latent TB infection (LTBI) acquired years earlier, or recent transmission. From a public health standpoint, recent transmission of TB is of particular concern as it has to potential to generate explosive outbreaks, particularly in vulnerable populations.[2–4] These outbreaks may fuel larger epidemics, leading to additional cases in the local community and fueling secondary outbreaks in other populations. Unfortunately, given the natural history of TB and lack of diagnostics to distinguish reactivation of LTBI and recent transmission, identifying who infected whom among active cases is notoriously challenging. There is some evidence that the majority of recent transmission is a result of "superspreading," an extreme manifestation of

inter-individual heterogeneity wherein a disproportionately small number of individuals contribute to the majority of secondary cases. However, major gaps remain in our understanding of inter-individual transmission and its importance in TB transmission dynamics.[5–8] As a result, researchers have explicitly called for an improved understanding of TB transmission dynamics, particularly the development of new methods that advance our understanding of superspreading in TB.[6–8]

As a partial answer to this call, this dissertation's overall goal is to quantify the degree of inter-individual heterogeneity in TB transmission. Quantifying such heterogeneity affords surveillance systems the ability to evaluate the propensity of superspreading in a population, improves modeling efforts through a more accurate representation of transmission heterogeneities, and informs targeted public health interventions. This goal is confronted by a well-known limitation in TB: the number of secondary cases arising from each infectious case is rarely, if ever, identified with any accuracy. Thus, researchers are unable to reliably quantify differences in secondary cases between infectious individuals. However, advancements in genotypic techniques combined with traditional epidemiologic approaches have allowed for the accurate identification of entire TB transmission clusters (defined as an index case and all subsequent cases arising from the index case). Leveraging the properties of transmission cluster distributions, this dissertation's goal is accomplished by: 1) developing and evaluating a method to reliably infer inter-individual heterogeneity using the distribution of final transmission cluster sizes in a surveillance system, 2) applying this method to transmission cluster data that was systematically abstracted from global surveillance systems in the published literature to quantify individual heterogeneity in various global populations, and 3) applying this method to data obtained from the Centers for Disease Control and Prevention (CDC) to assess individual heterogeneity and its role in transmission dynamics in the United States.

1.1.1 Study 1: Specific Aims and Summary

The specific aims of Study 1 were to:

1. To develop a method to estimate the degree of inter-individual heterogeneity in TB transmission using transmission cluster surveillance data

2. To evaluate the bias introduced by common limitations in TB surveillance and cluster data, including censorship, overlapping transmission clusters, and imperfect case ascertainment

3. To apply this method in estimating inter-individual heterogeneity in the United States using CDC-defined transmission cluster data

Briefly, utilizing a branching process model with a negative binomial offspring distribution, mechanistic adjustments were made to the probability distribution to relate the distribution of secondary cases to the distribution of final cluster sizes. Maximum likelihood estimation (MLE) was applied to infer both the reproductive number, R, and the dispersion parameter k. The dispersion parameter k of the negative binomial specifically quantifies individual heterogeneity. Simulations were used to compare the performance of the cluster-based inference procedure to full individual data to assess robustness under perfect surveillance. Adjustments were made to the simulation procedure to emulate common limitations with TB surveillance that affect cluster size data: imperfect case ascertainment (through both passive and active case ascertainment), censorship due to the sampling time frame, and overlapping clusters (wherein two chains of transmission cannot be unambiguously separated). The extent and direction of bias introduced by these limitations was first assessed univariately. Subsequently, three combined scenarios (emulating high-resource, moderate-resource, and low-resource settings) were constructed to assess multivariate bias. Finally, the epidemiologic utility of this method was evaluated by applying this method to transmission cluster data in the United States.

1.1.2 Study 2: Specific Aims and Summary

Study 2 applied the methods developed in Study 1 to quantify inter-individual heterogeneity in surveillance system data abstracted from the literature. Once heterogeneity was quantified by virtue of the parameter k, the study further applied the transmission parameter estimates to investigate the role that such heterogeneity plays in the broader context of transmission dynamics. The specific aims of study 2 were:

1. To quantify heterogeneity in TB transmission across global surveillance systems
    (a) To build a preliminary evidence base regarding the degree of heterogeneity present in various global contexts
2. To evaluate the impact of inter-individual heterogeneity on TB transmission dynamics

Briefly, we systematically gathered empirical TB transmission cluster size data from detailed contact tracing, whole genome sequencing, and epidemiological surveillance of TB transmission. $\hat{R}$ and $\hat{k}$ were jointly estimated using the methods defined in Study 1 to examine the extent of individual variation in secondary cases. To investigate the impact such variation may have on epidemic spread, we calculated the absolute and relative probability that a cluster initiating with a single index case would result in the largest cluster observed in the dataset under three distributional assumptions common in epidemiologic modeling: the negative binomial ($Y \sim NB(\hat{R}, \hat{k})$), geometric ($Y \sim GEO(\hat{R})$), and Poisson ($Y \sim POI(\hat{R})$).

1.1.3 Study 3: Specific Aims and Summary

Working with the United States Center for Disease Control and Prevention (U.S. CDC), Study 3 applied the methods developed in Study 1 to evaluate individual transmission heterogeneity in the four U.S. states with the highest number of incident TB cases: California, Florida, New York, and Texas. Using negative binomial

parameters $\hat{R}$ and $\hat{k}$ inferred from cluster-level data in these states, Study 3 further characterized TB transmission by quantifying the proportion of infectious cases responsible for a given proportion of transmission. The specific aims of Study 3 were:

1. To quantify heterogeneity in TB transmission in four U.S. states
2. To estimate the proportion infectious cases for a given percentage of secondary cases

Briefly, we used routinely collected data from the U.S. Centers for Disease Control and Prevention (CDC) National Tuberculosis Surveillance System (NTSS), the National Tuberculosis Genotyping Service (NTGS), and the Large Outbreaks of Tuberculosis in the United States (LOTUS) databases from January 1, 2014 to December 31, 2016 for the states of California, Florida, New York, and Texas. Transmission clusters were defined using 24-locus mycobacterial interspersed repetitive unit variable number of tandem repeats (MIRU-24) and whole genome sequencing (WGS). Transmission clusters were defined as cases with identical MIRU-24 profiles within the same county during the study timeframe. Clusters meeting LOTUS criteria were further evaluated using the more discriminatory WGS. Taking $\hat{R}$ and $\hat{k}$ to specify the exact probability density function (PDF) and cumulative density function for the given populations, we then calculated the expected proportion of transmission attributed to a specified proportion of the cases, p_t. For clarity, common example known to sexually transmitted and vector-borne diseases is the "80/20 rule," wherein 80 percent of transmission is attributable to only 20 percent of infectious cases (i.e. $p_t = 0.80$).

1.2 Tuberculosis Epidemiology and Transmission Dynamics

1.2.1 Global Tuberculosis Epidemiology

Tuberculosis disease (TB) is a directly transmitted infectious disease caused by the bacterial *Mycobacterium tuberculosis* (*Mtb*) that remains a major global health crisis. In 2018, there were an estimated 10 million (range: 9.0-11.1 million) incident cases of active TB disease and 1.5 (range: 1.3-1.6 million) million deaths worldwide.[1] The global burden of TB has been relatively stable over the past 20 years with a slight peak in the mid- to late-2000's followed by a subtle decline over the past decade (Figure 1.1). The incidence rate of TB has dropped 14 percent from 2000-2010 (median: 51 cases per 100,000 population) to 2010-2018 (median: 44 cases per 100,000 population). The rate of this decline has slowed, however, as more recently the incidence rate has only decreased 2 percent from 2012 to 2018.

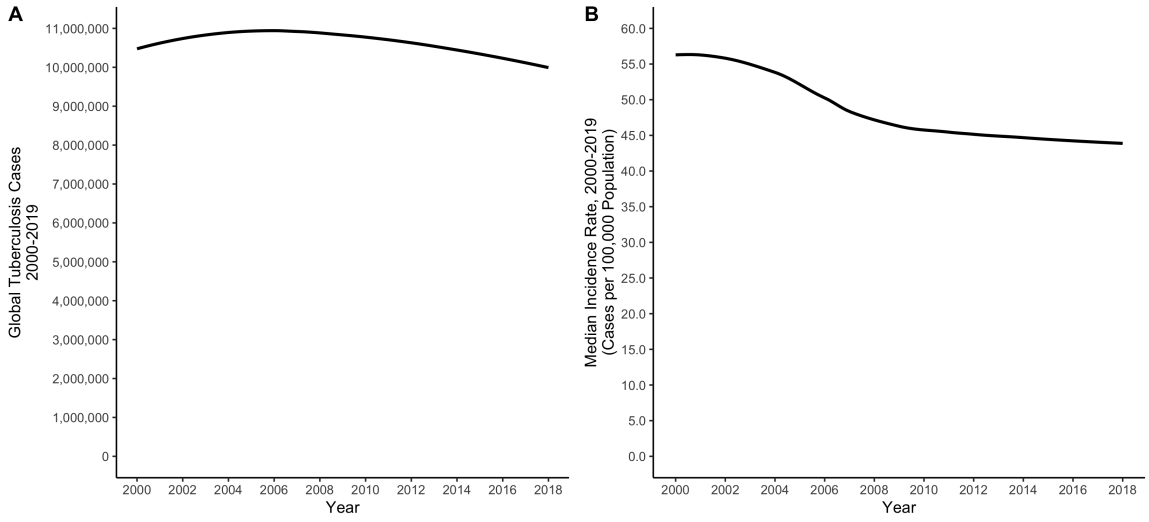

Figure 1.1: Global cases of tuberculosis, 2000-2018. A) Total TB cases worldwide, 2000-2018. B) Median global TB incident rate, per 100,000 population. Data from World Health Organization (2018).

Although the global epidemiology of TB suggests an improvement in TB control,

such aggregate data mask a markedly heterogeneous epidemic that varies widely between countries, places, and people. Many low- and middle-income countries (LMICs) experience incidence rates over 50 times that in most high-income countries, and two-thirds of all reported incident TB cases are found in only 8 countries: India (27%), China (9%), Indonesia (8%), the Philippines (6%), Pakistan (6%), Nigeria (4%), Bangladesh (4%), and South Africa (3%) (Figure 1.2, Panels A and B).[1,9] The consequences of these disparities are realized when examining death rates. TB-related deaths are significantly lower in high-income countries than in LMICs, and this reduction is disproportionate to the country's incidence: in general TB patients are more likely to die in LMIC countries than in high-income countries (Figure 1.2, Panels C and D). Such variability in these data suggest that dramatic reductions in TB morbidity and mortality are achievable by identifying modifiable prevention factors that determine transmission.[10–12]

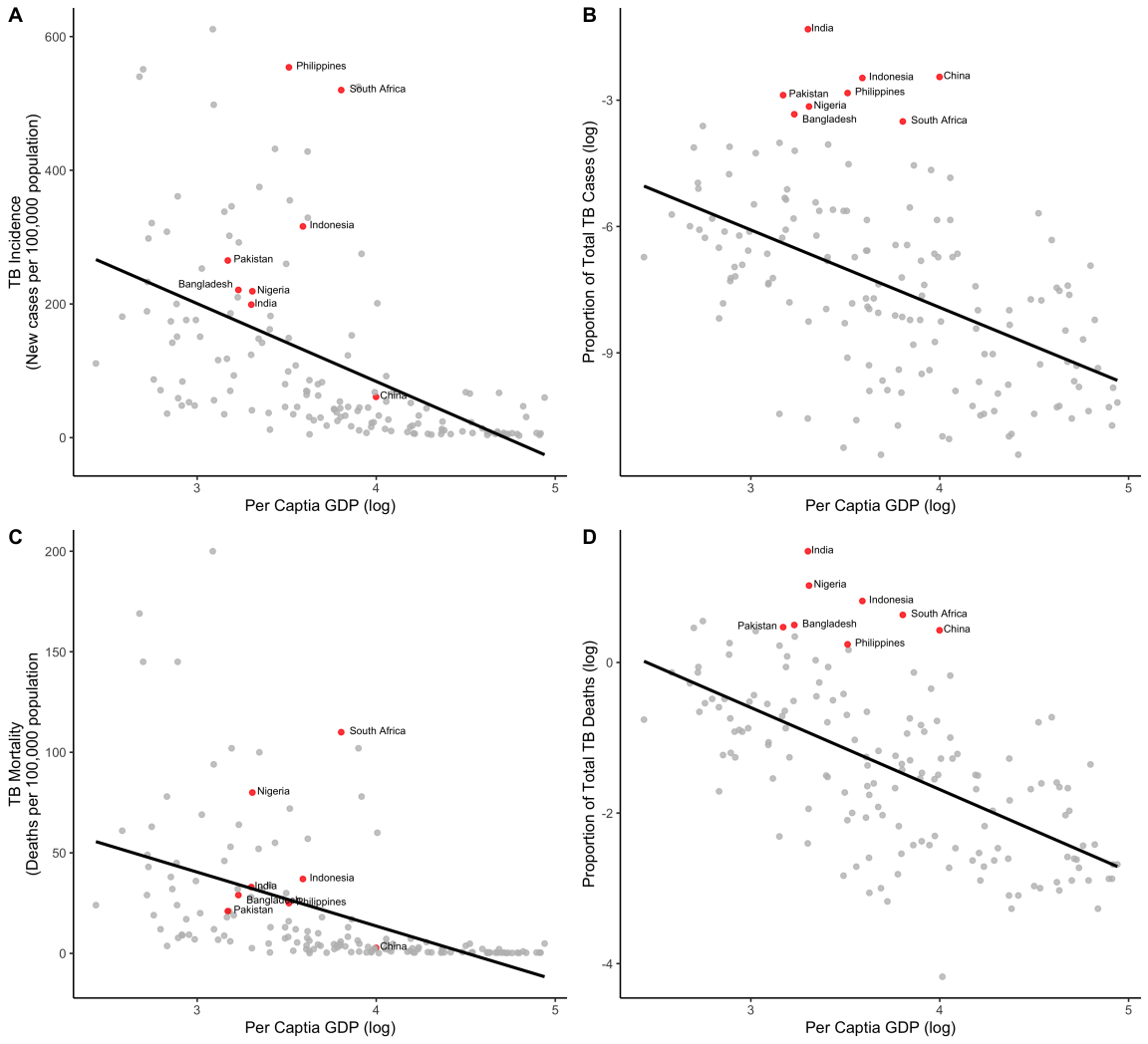

Figure 1.2: TB morbidity and mortality are related to country income. A) TB incidence rate (cases per 100,000 population) vs per capita GDP; B) Proportion of total TB cases vs per capita GDP; C) TB mortality rate (deaths per 100,000 population) vs per capita GDP; and D) Proportion of total TB deaths vs per capita GDP. Note the log scale for GDP (all figures) and proportions (B and D). The eight countries comprising 67 percent of the TB burden are highlighted in red. Data are from the World Bank and World Health Organization and include 156 countries with both GDP and TB incidence data (2018).

1.2.2 Inter-Individual Heterogeneity and Superspreading in Tuberculosis Transmission

In a given population, TB epidemiology is ultimately determined by individuals and their capacity to transmit and develop TB. Inter-individual differences in these capacities have been shown to greatly influence both the success of localized outbreaks and outbreak trajectory.[13] In this body of work, inter-individual heterogeneity in TB transmission is explicitly defined as differences in the number of secondary cases arising between infectious individuals within a population. The advantage of this definition is that it incorporates all known and unknown host, contact, pathogen, and environmental factors that lead to the number of secondary cases from an infectious individual (the "infectious history"). Importantly, although no universally accepted definition exists, "superspreading" is generally acknowledged as an extreme manifestation of inter-individual heterogeneity wherein few infectious individuals result in a disproportionate number of secondary cases, while the majority of cases lead to little or no ongoing transmission.

As evidenced from other infectious diseases, inter-individual heterogeneity and superspreading have a profound impact on both disease emergence and subsequent outbreak trajectory within a population.[14–17] The probability that an outbreak will emerge in a population after the introduction of an infectious case decreases as inter-individual heterogeneity increases (i.e. heterogeneity favors more rapid stochastic extinction in the early generations of spread).[14] This is largely due to the increased probability that a given index case will transmit few or no secondary cases as inter-individual heterogeneity increases. Conversely, the few populations that evade extinction after disease introduction are characterized by less predictable, more explosive outbreaks.[13] Such dynamics can be observed in a number of outbreaks over the past several decades, including the 2003 Severe Acute Respiratory Syndrome (SARS) outbreaks, the 2012 Middle East Respiratory Syndrome (MERS), and the 2014 Ebola

outbreak in West Africa. In these global outbreaks, the majority of communities experienced little to no epidemic spread after initial introduction of a source case, whereas others suffered large and explosive outbreaks.[18–20] A retrospective evaluation of these data show that the differences in these transmission patterns were largely accounted for by the presence or absence of a superspreader in the early generations of spread.[17,21,22]

Although little epidemiological research has investigated superspreading in TB transmission, similar patterns are observed in surveillance data and early evidence suggests that a high degree of inter-individual heterogeneity may be a defining feature shaping the epidemiology of TB.[5,6,8,11,16,23] For instance, Gardy *et al* used whole-genome sequencing and social-network analysis to describe a TB outbreak in British Columbia, Canada and found that individual superspreaders were the largest single contributing factor to overall TB prevalence in the study population.[7] Additionally, a recent modeling study investigating the location of TB transmission used empirical data from South Africa to incorporate superspreading into their mathematical framework.[8] The authors found that as the propensity for superspreading was reduced, a greater proportion of TB transmission was attributed to the household than to the general community. Another study using high-resolution follow up and contact tracing data over a ten-year period in Victoria, Australia concluded that superspreading is responsible for the majority of secondary cases in the community.[6] Such studies underscore the importance of characterizing and quantifying the degree of inter-individual heterogeneity in TB transmission in order to optimize public health interventions designed to interrupt transmission.[24]

1.2.3 Sources of Inter-Individual Heterogeneity in TB Transmission

For most infectious diseases, heterogeneity in secondary infections among infectious individuals is synonymous with heterogeneity in secondary cases.[13,16,25,26] However, TB transmission dynamics are uniquely complicated by latent TB infection (LTBI). LTBI is successful infection of *Mtb* in a susceptible host characterized by a dynamic and sustained balance of immune response and bacterial persistence that may last for decades; LTBI is not considered clinical TB disease and is non-infectious (Figure 1.3).[27] Thus, successful transmission and infection of *Mtb* may or may not result in a secondary case of TB; an estimated 90-95 percent of latently infected individuals never progress to active TB disease.[1] Reactivation of LTBI, wherein LTBI progresses to active TB disease, is highest in the first 2-5 years following infection yet may occur in the distant future due to changes in the host's immunological status.[28] This pathological hallmark of TB infection implies that while increased infectiousness of a source case may lead to an increased number of infections, the resultant number of observed secondary cases is critically dependent on factors beyond the host, namely the propensity and timing of the secondary host to progress to active disease after successful infection.[29]

The relationship between the number of secondary transmission events resulting from a source case and the number of observed secondary cases attributed to that case is unclear. Immunological, diagnostic, and practical challenges associated with LTBI detection thwart our ability to draw conclusions and make acquisition of the high-resolution data needed to determine when, how, and where TB infection occurred difficult. However, a recent well-designed study using detailed contact tracing and surveillance data over a 10-year period compared the distribution of secondary infections to the distribution of observed secondary cases.[6] The study found that, while the the magnitude of the distributions differed (by default there were more

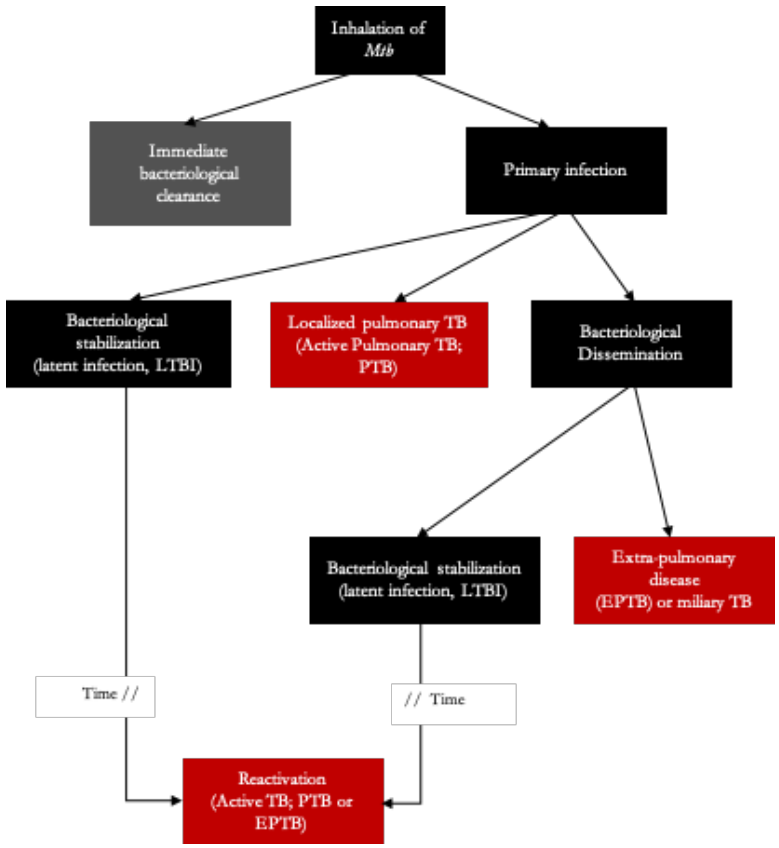

Figure 1.3: Pathways of disease occurrence after inhalation of *Mtb*

secondary infections than secondary cases), the overall shapes of the two distributions were similar. This implies that those who infected more individuals were responsible for more secondary cases. While this study provides preliminary evidence that individual infectiousness may be a correlate with the number of observed secondary cases, our understanding of this relationship is far from complete.

In contrast to the slow progression of LTBI to active disease, incident TB cases may also arise due to recent transmission. Recent transmission is distinguished from reactivation of LTBI as it focuses on the proportion of infected individuals who progress to active TB within a finite timeframe after infection (0-3 years[30]). From a public health standpoint, recent transmission is of particular importance as it represents the increased potential of rapid, extensive outbreaks, particularly in vulnerable communities.[3,13,31] Both recent transmission and LTBI are a consequence of myriad

known and unknown host, contact, pathogen, and environmental factors.

Characterizing heterogeneities within these sources to better understand how and where TB spreads is a basic public health principal. When considering differences in the number of observed cases due to recent transmission, it is useful to center one's thinking around three primary categories: 1) the infectivity of the source case; 2) the intensity of exposure to the susceptible individual; and 3) the susceptibility of the exposed person to infection and disease. Using this framework, this section briefly describes the most prominent sources of known heterogeneities in TB epidemiology.

Infectiousness of the source case

Here, infectiousness is defined as the capacity of an individual to transmit *Mtb* to susceptible contacts, including transmission that results in recent transmission or LTBI. Variation in the infectiousness of a source case is predicated on several known properties of disease manifestation. As TB is airborne and exposure to *Mtb* from an infectious individual is a necessary cause for infection, it is intuitive that one's ability to emit infectious droplet nuclei is correlated with a higher degree of infectiousness. Accordingly, extrapulmonary cases of TB (EPTB), such as TB of the bones or joints, are generally considered non-infectious unless the extrapulmonary disease is located along the breathing pathway (i.e. oral cavity, throat, or larynx).[32] Early studies investigating mechanics of pulmonary TB infection showed that patients with more severe manifestations of active disease may generate a larger quantity of more infectious droplets.[33,34] Along this vein, cough is the most common symptom of TB and the severity of disease may increase the frequency and strength of coughing.[35] Smear-positive adults (in which *Mtb* is positively identified under a microscope) and TB cases characterized by cavitary lung lesions have also been shown to transmit more extensively.[36] Conversely, young children are typically paucibacillary and rarely develop severe disease as compared to adults.[37]

Social, personal, and demographic factors also play an important role in determining one's infectiousness. While men account for a higher proportion of incident TB, female cases are typically younger, have lower rates of cavitation, and improved success once in care.[1,38,39] However, it is not well established if these gender differences are a product of biological mechanisms or if they reflect sociocultural norms influencing health behavior.[40–42] There is evidence that TB patients coinfected with HIV may transmit less than TB patient uninfected with HIV.[43,44] Though inconclusive, this hypothesis is in part because HIV/TB coinfection hastens the time to active TB disease and often results in more severe cases of TB, both of which reduce the duration of infectiousness (either through social isolation, faster access to necessary health services, or death).[45,46] HIV-infected patients are also less likely to be smear positive than HIV-uninfected patients.[47–49] However, with widespread and increasing use of ART in recent years, the dynamics of TB/HIV coinfection may change in the near future.

Mixing patterns and number of contacts also plays a role in the number of secondary infections. Certain occupations (i.e. restaurant server, teacher) may involve a high number of contacts, whereas others (i.e. PhD student) are more solitary. Distal or ecological factors, such as a community's public transport system and homelessness, also greatly impact contact rates of an infectious individual.[50–52] While active cases of TB can quickly become non-infectious with proper treatment, individuals differ in their desire or ability to seek medical care and the quality of care recieved.[53–55] Healthcare capacity also interplays with the strain of *Mtb* infecting the source case; while there is no consensus that drug-resistant TB (DR-TB) strains transmit more or less efficiently and drug susceptible strains, DR-TB strains are definitively more challenging and costly to treat and have considerably lower treatment success rates, all of which may increase an individual's ability to infect additional contacts.[1,56–58]

Intensity of exposure

Intensity of exposure refers to how long a susceptible contact is exposed to *Mtb* and the bacterial load to which they are exposed. Physical environments that increase or decrease duration and exposure to *Mtb* may play a key role in increased secondary cases. Prisons and jails are prominent examples, as they have been shown to be a key reservoir for TB transmission and a critical link to the general population.[59,60] TB is considered an occupational disease as certain occupations are prone to involve more contacts in prolonged and close proximity to infectious cases.[61] For instance, infection pressure is so great within South African gold mines that mass screening and treatment for LTBI had no effect on reducing secondary cases of TB.[62] Health-care workers are well-known to be at increased risk of *Mtb* infection, as they may be exposed to multiple infectious TB patients, work in poorly ventilated spaces, and perform procedures with contaminated aerosols.[63,64] Numerous additional examples exist of outbreaks occurring in other settings conducive to transmission, such as public transport, churches, informal social establishments (shebeens), and other social mixing patterns.[50,65–69]

Arguably the most critical aspect of intensity of exposure is within the household. Household contacts of an infectious case of TB are at substantially higher risk of TB infection and active TB disease.[70] However, the role of household contacts as it pertains to differences in the number of secondary cases arising from the index case is uncertain. While the absolute risk is indeed greater among household contacts, several studies have found that a larger proportion of secondary cases arise outside the household, and thus community transmission may be more influential in inter-individual heterogenenity.[43,66,71,72] This may be due to a phenomenon of "contact saturation;" additional exposures in the community are likely to naïve community members, whereas additional exposures among household members are "wasted" once the contact is infected with *Mtb*.[8,73]

Susceptibility of the exposed person

Here, susceptibility refers to both the likelihood of infection of a susceptible contact after exposure and the progress from *Mtb* infection to clinical disease after infection. Numerous studies have identified prominent comorbidities that increase susceptibility. HIV infection is the most well-described risk factor, with numerous studies demonstrating that HIV infection both increases the probability of infection and hastens the time to active TB disease.[52,74–76] For instance, a prospective cohort study found that after two years of follow up, the odds of developing active TB disease in HIV-infected individuals was 18.8 (95% Confidence Interval (CI): 10.3-34.5) higher than that of HIV-uninfected individuals, with an incidence of 25.3 and 1.3 per 1,000 person-years, respectively, for culture-positive active TB.[77] Silicosis, a form of pulmonary fibrosis resulting from inhalation of silica dust, is also known to increase the risk of active TB two- to three-fold compared to the general population.[78,79] Additional comorbidities are associated with substantially increased susceptibility, including those with chronic renal failure, diabetes mellitus, or those on corticosteroids and other immunosuppressive drugs.[80–84] Malnutrition profoundly compromises one's immune function and both increases the probability of infection and the timing and severity of clinical disease.[85,86] Alcohol use has also been shown to substantially impact both the risk and progression of TB disease.[87] LTBI itself may play a role in susceptibility of reinfection; there is some evidence to show that latently infected individuals have partial immunity to reinfection from subsequent exogenous exposures to *Mtb* and also have a lower risk of progressing to active disease if infection occurs.[88] Conversely, patients with previous episodes of active TB disease are more susceptible to reinfection despite treatment outcome.[89–91]

1.3 Current Approaches to Addressing Heterogeneity in TB Transmission

Heterogeneity in TB and its effect on transmission dynamics is most often explored through mathematical modeling. Mathematical models are often used in epidemiology to better understand epidemic spread, identify gaps in data needs, and predict how well interventions may work (either alone or in combination) to mitigate disease spread. Quantifying heterogeneities in transmission is a key component in epidemiologic modeling and is critical in evaluating the success of interventions.

Since the first TB model in 1962, hundreds of studies have investigated diverse aspects of TB transmission dynamics using various approaches to mathematical modelling.[92] Due to a combination of computational convenience and accuracy, the vast majority of TB models are dynamic, compartmental, deterministic transmission models (Table 1.1).[93] While the model structures themselves are diverse depending on what scientific question the study is trying to answer, all deterministic compartmental models account for heterogeneity by structuring the population of interest into distinct sub-groups based on known factors associated with transmission; each group is itself assumed to be homogenous. These models often perform well in their estimation and account for many of the sources of heterogeneity in TB transmission. For example, variable infectiousness is typically represented by smear status and susceptibility is often accounted for by groups representing "fast" and "slow" progression to active disease.[94,95]

Table 1.1: Brief summary of epidemiologic models used in assessing infectious disease transmission dynamics

Model Types	Description
Dynamic vs Static Models	Dynamic models account for time-dependent changes in model parameters to model risk of infection; incidence is a function of prevalence. Static models are not sensitive to time-dependent factors.
Compartmental vs Individual-Based Models	In compartmental models, disease and immunity states are modeled at the group level. Individual-based models (also known as 'agent-based' models) use individuals as the level of analysis
Stochastic vs Deterministic Models	Stochastic models account for inherent randomness in transmission in addition to other model inputs. In deterministic models, the model output is wholly determined by input parameter values and conditions.

Numerous well-designed deterministic compartmental models have proved integral to our understanding of TB transmission and prevention.[96–98] However, an important limitation as it pertains to evaluating inter-individual heterogeneity in TB transmission is that groups must be identified according to some known, identifiable property (as described above). This presents obvious challenges to incorporating superspreading in TB transmission, which often cannot be predicted *a priori*. For instance, Curtis *et al* describe widespread TB transmission from a nine-year-old boy with extensive bilateral cavitary TB; meanwhile his identical twin brother experienced a relatively mild case and was not considered infectious.[99] Additionally, Edwards *et al* demonstrate significant variation among individuals' ability to emit exhaled bioaerosols gen-

erated during normal breathing; identifying two distinct groups ("high-producers" and "low-producers" with averages of 1580 and 38 particles per liter, respectively) with no discernible characteristic differences between the two.[100] Given the nuance in individual differences, deterministic compartmental models are unable to capture the full extent of individual heterogeneity in TB transmission.

Recent advancements in computer power have afforded the increasing use of individual-based models (also known as agent-based models) to study the transmission of TB. The fundamental units of analysis in these models are individuals; typically, the model specifies a distribution of discrete individuals, with each individual assigned a vector of values that modulate risk of infection, progression to clinical disease, and infectiousness.[101,102] In addition, several models assign individuals to demographic or physical spaces, such as households or neighborhoods.[8,101–103] A subset of individual-based models are known as network models, in which edges of a graph depict relationships among individuals in a population (nodes).[104,105] These models further complement analyses by considering the importance of structure, pathways of infection, and social networks. Network models often represent individual heterogeneity using degree distributions (i.e. the number of connections one node has to other nodes).

While using the same deterministic skeleton of traditional compartmental models, individual-based models are useful in the context of understanding the importance of inter-individual heterogeneity of TB transmission. These models track the disease and immunity states of individuals, and thus afford the opportunity to assign individual-level mechanistic and stochastic processes associated with transmission. As a result, individual-based models are required to account for unknown individual variation that may lead to superspreading. However, their use is far less widespread than compartmental models; in contrast to the hundreds of compartmental models, from 2005 to 2016 only 26 studies using individual based models were published, with

only 18 of them specifically focusing on transmission dynamics and interventions.[106] To date only one model investigating TB transmission dynamics has specifically incorporated the propensity for superspreading in a population into the model as separate parameters.[8] In line with models of other infectious diseases, this added mechanism to the model accounted for all unknown factors associated with infectiousness and susceptibility. Their results more accurately recreated observed patterns of transmission and concluded that that many of the unexplained phenomena observed in TB transmission patterns may be due to the presence of superspreading. However, while the authors used empirical data to estimate these model parameters, the methods used for its estimation were limited and the data were incomplete (only 27 percent of participants had available data for parameter estimation). The authors conclude that an improved understanding of superspreading and novel methods to quantify superspreading in TB transmission are paramount to improved intervention strategies.

Although current modeling approaches often account for known factors contributing to such heterogeneity, unknown factors of (and the unknown interactions between) the host, contact, environment, and bacilli make the accurate representation of inter-individual heterogeneity one of the great ongoing challenges in TB modeling. This dissertation addresses this ambitious challenge, in part, by first developing a method to infer the propensity for superspreading in TB transmission. By doing so, stochastic individual-based models may be used to incorporate a separate model parameter accounting for all unknown factors attributed to inter-individual heterogeneity. We then apply this method to accurately establish empirical estimates of this parameter across various global contexts. By accounting for unknown factors attributed to individual variation, future modeling efforts can provide more accurate model predictions of transmission dynamics and improving the evaluation of interventions.

1.4 Chapter 1 References

1. Global Tuberculosis Report 2019. Geneva: World Health Organization; 2019.

2. Makhado NA, Matabane E, Faccin M, et al. Outbreak of multidrug-resistant tuberculosis in South Africa undetected by WHO-endorsed commercial tests: an observational study. The Lancet Infectious Diseases 2018;18:1350-9.

3. Haddad M, B. , Mitruka M, B. , Oeltmann J, Johnson WD, Jr., Navin TR. Characteristics of Tuberculosis Cases that Started Outbreaks in the United States, 2002–2011. Emerging Infectious Disease 2015;21:508.

4. Norheim G, Seterelv S, Arnesen TM, et al. Tuberculosis Outbreak in an Educational Institution in Norway. Journal of clinical microbiology 2017;55:1327-33.

5. Ypma RJ, Altes HK, van Soolingen D, Wallinga J, van Ballegooijen WM. A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. Epidemiology (Cambridge, Mass) 2013;24:395-400.

6. Melsew YA, Gambhir M, Cheng AC, et al. The role of super-spreading events in Mycobacterium tuberculosis transmission: evidence from contact tracing. BMC Infectious Diseases 2019;19:244.

7. Gardy JL, Johnston JC, Sui SJH, et al. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. New England Journal of Medicine 2011;364:730-9.

8. McCreesh N, White RG. An explanation for the low proportion of tuberculosis that results from transmission between household and known social contacts. Scientific Reports 2018;8:5382.

9. World Bank Country and Lending Groups. Country Classification. http://databank.worldbank.o
content/CLASS.xls: World Bank; 2020.

10. Dowdy DW, Azman AS, Kendall EA, Mathema B. Transforming the Fight Against Tuberculosis: Targeting Catalysts of Transmission. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America 2014;59:1123-

9.

11. Trauer JM, Dodd PJ, Gomes MGM, et al. The Importance of Heterogeneity to the Epidemiology of Tuberculosis. Clinical Infectious Diseases 2018;69:159-66.

12. Lönnroth K, Jaramillo E, Williams BG, Dye C, Raviglione M. Drivers of tuberculosis epidemics: The role of risk factors and social determinants. Social Science and Medicine 2009;68:2240-6.

13. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438:355-9.

14. Lipsitch M, Cohen T, Cooper B, et al. Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. Science (New York, NY) 2003;300:1966-70.

15. Shen Z, Ning F, Zhou W, et al. Superspreading SARS Events, Beijing, 2003. Emerging Infectious Diseases 2004;10:256-60.

16. Stein RA. Super-spreaders in infectious diseases. Int J Infect Dis 2011;15:e510-3.

17. Wong G, Liu W, Liu Y, Zhou B, Bi Y, Gao George F. MERS, SARS, and Ebola: The Role of Super-Spreaders in Infectious Disease. Cell Host and Microbe 2015;18:398-401.

18. Severe acute respiratory syndrome–Singapore, 2003. MMWR Morbidity and mortality weekly report 2003;52:405-11.

19. Anderson RM, Fraser C, Ghani AC, et al. Epidemiology, transmission dynamics and control of SARS: the 2002-2003 epidemic. Philosophical transactions of the Royal Society of London Series B, Biological sciences 2004;359:1091-105.

20. Breban R, Riou J, Fontanet A. Interhuman transmissibility of Middle East respiratory syndrome coronavirus: estimation of pandemic risk. Lancet (London, England) 2013;382:694-9.

21. Alhamlan FS, Majumder MS, Brownstein JS, et al. Case characteristics among Middle East respiratory syndrome coronavirus outbreak and non-outbreak cases in Saudi Arabia from 2012 to 2015. BMJ Open 2017;7:e011865.

22. Kucharski AJ, Althaus CL. The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. Eurosurveillance 2015;20:21167.

23. Mitruka K, Oeltmann J, Ijaz K, Haddad M. Tuberculosis Outbreak Investigations in the United States, 2002–2008. Emerging Infectious Disease Journal 2011;17:425.

24. Dye C, Williams BG. The Population Dynamics and Control of Tuberculosis. Science (New York, NY) 2010;328:856-61.

25. Blumberg S, Lloyd-Smith JO. Comparing methods for estimating R0 from the size distribution of subcritical transmission chains. Epidemics 2013;5:131-45.

26. Getz WM, Lloyd-Smith JO, Cross PC, et al. Modeling the invasion and spread of contagious diseases in heterogeneous populations. Disease Evolution: Models, Concepts, and Data Analyses; 2006.

27. Kaufmann SHE. How can immunology contribute to the control of tuberculosis? Nature Reviews Immunology 2001;1:20.

28. Comstock GW. Epidemiology of tuberculosis. Am Rev Respir Dis 1982;125:8-15.

29. Drain PK, Bajema KL, Dowdy D, et al. Incipient and Subclinical Tuberculosis: a Clinical Review of Early Stages and Progression of Infection. Clinical microbiology reviews 2018;31:e00021-18.

30. Schwartz NG, Price SF, Pratt RH, Langer AJ. Tuberculosis - United States, 2019. MMWR Morbidity and mortality weekly report 2020;69:286-9.

31. Yuen CM, Kammerer JS, Marks K, Navin TR, France AM. Recent Transmission of Tuberculosis — United States, 2011–2014. PLOS ONE 2016;11:e0153728.

32. Golden MP, Vikram HR. Extrapulmonary tuberculosis: an overview. Am Fam Physician 2005;72:1761-8.

33. Wells WF, Ratcliffe HL, Grumb C. On the mechanics of droplet nuclei infection; quantitative experimental air-borne tuberculosis in rabbits. American journal of hygiene 1948;47:11-28.

34. Riley RL, Wells WF, Mills CC, Nyka W, McLean RL. Air hygiene in tuberculosis:

quantitative studies of infectivity and control in a pilot ward. American review of tuberculosis 1957;75:420-31.

35. American Thoracic S. Diagnostic Standards and Classification of Tuberculosis in Adults and Children. American journal of respiratory and critical care medicine 2000;161:1376-95.

36. Melsew YA, Doan TN, Gambhir M, Cheng AC, McBryde E, Trauer JM. Risk factors for infectiousness of patients with tuberculosis: a systematic review and meta-analysis. Epidemiology and infection 2018;146:345-53.

37. Perez-Velez CM, Marais BJ. Tuberculosis in Children. New England Journal of Medicine 2012;367:348-61.

38. Murphy ME, Wills GH, Murthy S, et al. Gender differences in tuberculosis treatment outcomes: a post hoc analysis of the REMoxTB study. BMC Medicine 2018;16:189.

39. van den Hof S, Najlis C, Bloss E, Straetemans M. A systematic review on the role of gender in tuberculosis control. The Haugue, Netherlands: KNCV Tuberculosis Foundation; 2010.

40. Weiss MG, Sommerfeld J, Uplekar MW. Social and cultural dimensions of gender and tuberculosis. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease 2008;12:829-30.

41. Horton KC, MacPherson P, Houben RMGJ, White RG, Corbett EL. Sex Differences in Tuberculosis Burden and Notifications in Low- and Middle-Income Countries: A Systematic Review and Meta-analysis. PLOS Medicine 2016;13:e1002119.

42. Gosoniu GD, Ganapathy S, Kemp J, et al. Gender and socio-cultural determinants of delay to diagnosis of TB in Bangladesh, India and Malawi. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease 2008;12:848-55.

43. Middelkoop K, Mathema B, Myer L, et al. Transmission of tuberculosis in a South African community with a high prevalence of HIV infection. J Infect Dis 2015;211:53-61.

44. Espinal MA, Peréz EN, Baéz J, et al. Infectiousness of Mycobacterium tuberculosis in HIV-1-infected patients with tuberculosis: a prospective study. The Lancet 2000;355:275-80.

45. Chamie G, Luetkemeyer A, Charlebois E, Havlir DV. Tuberculosis as part of the natural history of HIV infection in developing countries. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 2010;50 Suppl 3:S245-S54.

46. Daley CL, Small PM, Schecter GF, et al. An outbreak of tuberculosis with accelerated progression among persons infected with the human immunodeficiency virus. An analysis using restriction-fragment-length polymorphisms. The New England journal of medicine 1992;326:231-5.

47. Lawn SD, Wood R. Tuberculosis in antiretroviral treatment services in resource-limited settings: addressing the challenges of screening and diagnosis. The Journal of infectious diseases 2011;204 Suppl 4:S1159-S67.

48. Wood R, Middelkoop K, Myer L, et al. Undiagnosed tuberculosis in a community with high HIV prevalence: implications for tuberculosis control. American journal of respiratory and critical care medicine 2007;175:87-93.

49. Davis JL, Worodria W, Kisembo H, et al. Clinical and radiographic factors do not accurately diagnose smear-negative tuberculosis in HIV-infected inpatients in Uganda: a cross-sectional study. PLoS One 2010;5:e9859.

50. Andrews JR, Morrow C, Wood R. Modeling the Role of Public Transportation in Sustaining Tuberculosis Transmission in South Africa. American journal of epidemiology 2013;177:556-61.

51. Connors WJ, Hussen SA, Holland DP, Mohamed O, Andes KL, Goswami ND.

Homeless shelter context and tuberculosis illness experiences during a large outbreak in Atlanta, Georgia. Public Health Action 2017;7:224-30.

52. Moss AR, Hahn JA, Tulsky JP, Daley CL, Small PM, Hopewell PC. Tuberculosis in the homeless. A prospective study. American journal of respiratory and critical care medicine 2000;162:460-4.

53. Charles N, Thomas B, Watson B, Raja Sakthivel M, Chandrasekeran V, Wares F. Care seeking behavior of chest symptomatics: a community based study done in South India after the implementation of the RNTCP. PLoS One 2010;5.

54. Gamtesa DF, Tola HH, Mehamed Z, Tesfaye E, Alemu A. Health care seeking behavior among presumptive tuberculosis patients in Ethiopia: a systematic review and meta-analysis. BMC Health Services Research 2020;20:445.

55. Cazabon D, Alsdurf H, Satyanarayana S, et al. Quality of tuberculosis care in high burden countries: the urgent need to address gaps in the care cascade. International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases 2017;56:111-6.

56. Bonnet M, Pardini M, Meacci F, et al. Treatment of Tuberculosis in a Region with High Drug Resistance: Outcomes, Drug Resistance Amplification and Re-Infection. PLOS ONE 2011;6:e23081.

57. Falzon D, Jaramillo E, Schünemann HJ, et al. WHO guidelines for the programmatic management of drug-resistant tuberculosis: 2011 update. European Respiratory Journal 2011;38:516.

58. Shrestha S, Knight GM, Fofana M, et al. Drivers and trajectories of resistance to new first-line drug regimens for tuberculosis. Open Forum Infect Dis 2014;1:ofu073-ofu.

59. Lambert LA, Armstrong LR, Lobato MN, Ho C, France AM, Haddad MB. Tuberculosis in Jails and Prisons: United States, 2002-2013. American journal of public health 2016;106:2231-7.

60. Sacchi FPC, Praça RM, Tatara MB, et al. Prisons as reservoir for community transmission of tuberculosis, Brazil. Emerging infectious diseases 2015;21:452-5.

61. Field M, ed. Tuberculosis in the Workplace. Washington, DC: United States Institute of Medicine Committee on Regulating Occupational Exposure to Tuberculosis; 2001.

62. Churchyard GJ, Fielding KL, Lewis JJ, et al. A Trial of Mass Isoniazid Preventive Therapy for Tuberculosis Control. New England Journal of Medicine 2014;370:301-10.

63. Joshi R, Reingold AL, Menzies D, Pai M. Tuberculosis among health-care workers in low- and middle-income countries: a systematic review. PLoS Med 2006;3:e494.

64. Menzies D, Joshi R, Pai M. Risk of tuberculosis infection and disease associated with work in health care settings. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease 2007;11:593-605.

65. Munch Z, Van Lill SW, Booysen CN, Zietsman HL, Enarson DA, Beyers N. Tuberculosis transmission patterns in a high-incidence area: a spatial analysis. The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease 2003;7:271-7.

66. Verver S, Warren RM, Munch Z, et al. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. Lancet (London, England) 2004;363:212-4.

67. Horna-Campos OJ, Sánchez-Pérez HJ, Sánchez I, Bedoya A, Martín M. Public Transportation and Pulmonary Tuberculosis, Lima, Peru. Emerging Infectious Diseases 2007;13:1491-3.

68. Johnstone-Robertson SP, Mark D, Morrow C, et al. Social mixing patterns within a South African township community: implications for respiratory disease transmission and control. American journal of epidemiology 2011;174:1246-55.

69. Chamie G, Wandera B, Marquez C, et al. Identifying locations of recent TB transmission in rural Uganda: a multidisciplinary approach. Tropical Medicine and International Health 2015;20:537-45.

70. Martinez L, Shen Y, Mupere E, Kizza A, Hill PC, Whalen CC. Transmission of Mycobacterium Tuberculosis in Households and the Community: A Systematic Review and Meta-Analysis. American journal of epidemiology 2017;185:1327-39.

71. Wilkinson D, Pillay M, Crump J, Lombard C, Davies GR, Sturm AW. Molecular epidemiology and transmission dynamics of Mycobacterium tuberculosis in rural Africa. Tropical medicine and international health : TM and IH 1997;2:747-53.

72. Glynn JR, Guerra-Assunção JA, Houben RMGJ, et al. Whole Genome Sequencing Shows a Low Proportion of Tuberculosis Disease Is Attributable to Known Close Contacts in Rural Malawi. PLOS ONE 2015;10:e0132840.

73. Eames KT. Modelling disease spread through random and regular contacts in clustered populations. Theor Popul Biol 2008;73:104-11.

74. Horsburgh CR, Rubin EJ. Latent Tuberculosis Infection in the United States. 2011;364:1441-8. 7

5. Selwyn PA, Hartel D, Lewis VA, et al. A Prospective Study of the Risk of Tuberculosis among Intravenous Drug Users with Human Immunodeficiency Virus Infection. New England Journal of Medicine 1989;320:545-50.

76. Wolday D, Hailu B, Girma M, Hailu E, Sanders E, Fontanet A L. Low CD4+ T-cell count and high HIV viral load precede the development of tuberculosis disease in a cohort of HIV-positive Ethiopians. The International Journal of Tuberculosis and Lung Disease 2003;7:110-6.

77. Corbett EL, Bandason T, Cheung YB, et al. Epidemiology of Tuberculosis in a High HIV Prevalence Population Provided with Enhanced Diagnosis of Symptomatic Disease. PLOS Medicine 2007;4:e22.

78. Rees D, Murray J. Silica, silicosis and tuberculosis. The international journal of

tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease 2007;11:474-84.

79. Cowie RL. The epidemiology of tuberculosis in gold miners with silicosis. American journal of respiratory and critical care medicine 1994;150:1460-2.

80. Hussein MM, Mooij JM, Roujouleh H. Tuberculosis and chronic renal disease. Semin Dial 2003;16:38-44.

81. Dooley KE, Chaisson RE. Tuberculosis and diabetes mellitus: convergence of two epidemics. The Lancet Infectious diseases 2009;9:737-46.

82. Cisneros JR, Murray KM. Corticosteroids in tuberculosis. Ann Pharmacother 1996;30:1298-303.

83. Gama L, Miranda C, Vargas R, et al. Immunosuppressant Drugs and Tuberculosis: Patient's Features in a University Hospital. C58 TUBERCULOSIS INFECTION AND DISEASE:A5524-A.

84. Skogberg K, Ruutu P, Tukiainen P, Valtonen V. Effect of Immunosuppressive Therapy on the Clinical Presentation and Outcome of Tuberculosis. Clinical Infectious Diseases 1993;17:1012-7.

85. Macallan DC. Malnutrition in tuberculosis. Diagn Microbiol Infect Dis 1999;34:153-7.

86. Hood MLH. A narrative review of recent progress in understanding the relationship between tuberculosis and protein energy malnutrition. European Journal of Clinical Nutrition 2013;67:1122-8.

87. Lönnroth K, Williams BG, Stadlin S, Jaramillo E, Dye C. Alcohol use as a risk factor for tuberculosis - a systematic review. BMC Public Health 2008;8:289.

88. Andrews JR, Noubary F, Walensky RP, Cerda R, Losina E, Horsburgh CR. Risk of progression to active tuberculosis following reinfection with Mycobacterium tuberculosis. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 2012;54:784-91.

89. Verver S, Warren RM, Beyers N, et al. Rate of Reinfection Tuberculosis after Successful Treatment Is Higher than Rate of New Tuberculosis. American journal of respiratory and critical care medicine 2005;171:1430-5.

90. Gomes MGM, Aguas R, Lopes JS, et al. How host heterogeneity governs tuberculosis reinfection? Proc Biol Sci 2012;279:2473-8.

91. Marx FM, Floyd S, Ayles H, Godfrey-Faussett P, Beyers N, Cohen T. High burden of prevalent tuberculosis among previously treated people in Southern Africa suggests potential for targeted control interventions. The European respiratory journal 2016;48:1227-30.

92. Waaler H, Geser A, Andersen S. The use of mathematical models in the study of the epidemiology of tuberculosis. Am J Public Health Nations Health 1962;52:1002-13.

93. Ozcaglar C, Shabbeer A, Vandenberg SL, Yener B, Bennett KP. Epidemiological models of Mycobacterium tuberculosis complex infections. Mathematical Biosciences 2012;236:77-96.

94. Pienaar E, Fluitt AM, Whitney SE, Freifeld AG, Viljoen HJ. A Model of Tuberculosis Transmission and Intervention Strategies in an Urban Residential Area. Computational biology and chemistry 2010;34:86-96.

95. Dye C, Williams BG. Eliminating human tuberculosis in the twenty-first century. J R Soc Interface 2008;5:653-62.

96. Lin H-H, Dowdy D, Dye C, Murray M, Cohen T. The impact of new tuberculosis diagnostics on transmission: why context matters. Bulletin of the World Health Organization 2012;90:739-47A.

97. Dowdy DW, Golub JE, Chaisson RE, Saraceni V. Heterogeneity in tuberculosis transmission and the role of geographic hotspots in propagating epidemics. Proceedings of the National Academy of Sciences of the United States of America 2012;109:9557-62.

98. Cohen T, Murray M. Modeling epidemics of multidrug-resistant M. tuberculosis of heterogeneous fitness. Nat Med 2004;10:1117-21.

99. Curtis AB, Ridzon R, Vogel R, et al. Extensive Transmission of Mycobacterium tuberculosis from a Child. New England Journal of Medicine 1999;341:1491-5.

100. Edwards DA, Man JC, Brand P, et al. Inhaling to mitigate exhaled bioaerosols. Proceedings of the National Academy of Sciences of the United States of America 2004;101:17383-8.

101. Shrestha S, Hill AN, Marks SM, Dowdy DW. Comparing Drivers and Dynamics of Tuberculosis in California, Florida, New York, and Texas. American journal of respiratory and critical care medicine 2017;196:1050-9.

102. Shrestha S, Cherng S, Hill AN, et al. Impact and Effectiveness of State-Level Tuberculosis Interventions in California, Florida, New York, and Texas: A Model-Based Analysis. American journal of epidemiology 2019;188:1733-41.

103. Murray M. Determinants of cluster distribution in the molecular epidemiology of tuberculosis. Proceedings of the National Academy of Sciences of the United States of America 2002;99:1538-43.

104. Cohen T, Colijn C, Finklea B, Murray M. Exogenous re-infection and the dynamics of tuberculosis epidemics: local effects in a network model of transmission. J R Soc Interface 2007;4:523-31.

105. Nelson KN, Gandhi NR, Mathema B, et al. Modeling Missing Cases and Transmission Links in Networks of Extensively Drug-Resistant Tuberculosis in KwaZulu-Natal, South Africa. American journal of epidemiology 2020;189:735-45.

106. Willem L, Verelst F, Bilcke J, Hens N, Beutels P. Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). BMC Infectious Diseases 2017;17:612.

Chapter 2

Quantifying Inter-Individual Heterogeneity in Tuberculosis Transmission

2.1 Background

The purpose of this chapter is to propose the overall theoretical framework used to quantify inter-individual heterogeneity in TB transmission and introduce the fundamental methods and terminology utilized in the process. Methods of quantifying inter-individual heterogeneity are relatively straightforward when using individual-level data: the number of secondary cases attributed to each infectious case are described in a probability distribution and the overdispersion of that distribution can be quantified. These methods cannot be applied to TB transmission data since LTBI and other issues prevent our ability to determine discrete transmission events and directionality. However, recent advancements in genetic techniques, such as whole genome sequencing (WGS), have been combined with traditional epidemiologic techniques (i.e. contact tracing) to reliably identify entire transmission clusters. Transmission

clusters are defined as an index case and all subsequent cases (i.e. secondary, tertiary, etc.) arising from that index case.

Much like the number of secondary cases from each infectious individual in a population follows a probability distribution, the collection of total transmission cluster sizes in the population also follows a distribution. While these two distributions are not the same, they are also not independent since the individual cases in each transmission chain give rise to the transmission cluster sizes in a given surveillance system (Figure 2.1). This dissertation seeks to exploit this relationship using branching process and probability theory to quantify inter-individual heterogeneity using only cluster-level data.

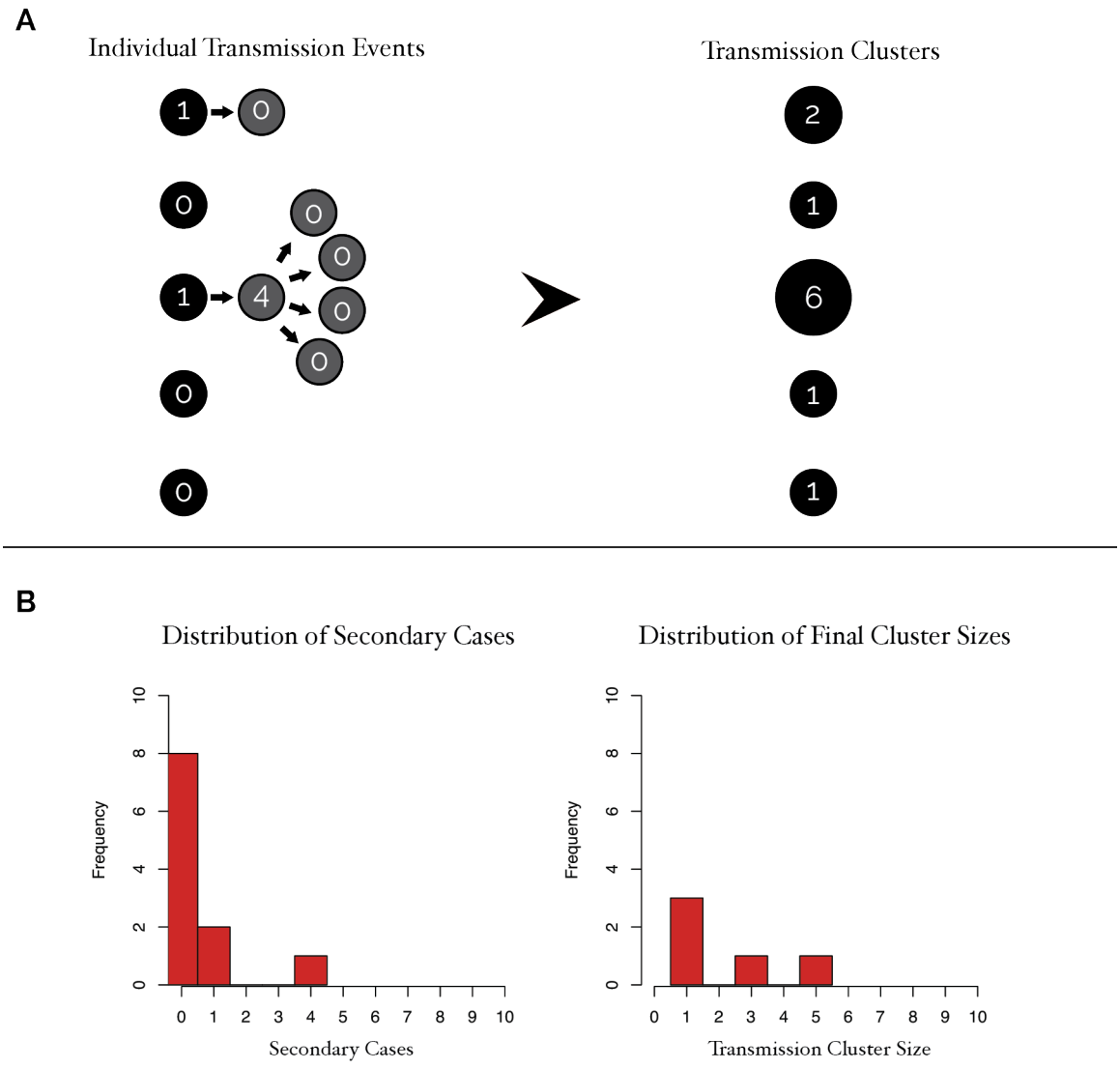

Figure 2.1: Visualization of individual- vs cluster-level data of TB transmission in a surveillance system. A) Individual-level data allow for the identification of transmission chains. Here, five independent index cases (black circles) are introduced to the population. Subsequent secondary transmission (grey circles) is represented horizontally. Individuals are represented as circles, and numerals indicate secondary cases attributed to each case. These chains of transmission give rise to the final transmission cluster sizes. B) Corresponding distributions of individual and cluster-level data for these example data.

2.2 Branching Process Analysis

2.2.1 Branching Process Overview

Branching processes are stochastic individual-based models that have long been used in ecology and epidemiology to model disease spread, and are of particular use in surveillance data as they only require information on the number of cases.[1,2] This analysis models transmission using a single-type branching process, also known as a discrete time Galton-Watson process. This is the most well-studied and validated approach to branching processes.[3] Galton-Watson branching processes assumes each infected individual is associated with a fixed length time interval known as a generation; at the end of each generation the individual will have produced a random number of secondary infections ("offspring"), herein denoted Z. Importantly, the basic reproductive number, R_0 (herein referred to as the more generalizable R), which represents the expected number of secondary cases caused by a typical infectious individual in a wholly susceptible population, is by definition the mean value of Z.

In branching processes of infectious disease, the offspring distribution is the probability distribution for the observed number of secondary cases caused by each individual infectious case, Z (i.e. $p_z = P(Z = z)$ for $z = 0, 1, 2, 3, \ldots, n$). Instrumental in the analysis of a branching process model is the probability generating function (pgf) of the offspring distribution. The pgf is a mathematical tool to study the sequence of probabilities and contains all the information needed to recover the probabilities associated with each Z value. The pgf in branching processes can generally be expressed as:[1,4]

$$G_z(s) = p_0 + p_1 s + p_2 s^2 + ... + p_n s^n = \sum_{z=0}^{\infty} p_z s^z \tag{2.1}$$

Where s is a dummy variable with no tangible meaning; the powers of s serve as a placeholder to recover the probabilities associated with Z and facilitate the use of high-order derivatives in their recovery.

2.2.2 Incorporating Inter-Individual Heterogeneity in Branching Process Models

Typically branching processes assume the number of cases resulting from an individual in each generation, Z, is a product of a Poisson process with intensity λ (i.e. $Z \sim POI(\lambda)$).[5] This allows branching process models to easily incorporate individual heterogeneity by allowing λ to itself be a random variable; each infectious case is associated with an individual reproductive number (denoted ν) drawn from some probability distribution representing the expected value of secondary cases for that specific individual (generally denoted here as $f_v(u)$).[6] The observed number of secondary cases, Z, is therefore a mixture of both ν and demographic stochasticity inherent in transmission. This approach can be generally represented by:[3,5]

$$G_z(s) = \int_0^\infty e^{-u(1-s)} f_v(u) du \tag{2.2}$$

In this context, the underlying epidemiologic mechanism of disease transmission is represented by virtue of the distributional assumption of λ and demographic stochasticity in disease transmission is modeled by the Poisson process. For instance, if there is no mechanistic plausibility for inter-individual heterogeneity (i.e. homogeneous transmission), $\nu = R$, and $Z \sim POI(R)$ thus all differences in observed Z values are a solely attributed to demographic stochasticity. Differential equation models often provide some mechanism for transmission dynamics that violate homogeneity. If it is assumed that ν is exponentially distributed with mean R, as is the case with many differential equation models with homogeneous transmission within groups and constant recovery rates, the resulting offspring distribution becomes geometrically distributed, $Z \sim GEO(R)$.

Both of these common distributional assumptions in epidemiology have a single parameter (R) and thus have an inherent and fixed assumption regarding inter-

individual heterogeneity. To allow for an unknown degree of inter-individual heterogeneity, we follow previous studies in assuming that ν is gamma distributed with mean R and dispersion parameter k.[6–9] This gamma-Poisson mixture results in a negative binomial offspring distribution of Z, also with mean R and dispersion parameter k (i.e. $Z \sim NB(R,k)$).[10] The dispersion parameter k quantifies the overdispersion of the distribution and encapsulates for the all known and unknown factors contributing to inter-individual heterogeneity in transmission.

Lemma 2.2.1. *A Poisson distribution with a gamma distributed λ with mean R and dispersion k results in a negative binomial distribution of Z, also with mean R and dispersion k:*

$$\begin{aligned} P(Z=z) &= \frac{1}{\Gamma(k)R^k}\int_0^\infty \frac{e^{-\lambda}\lambda^z}{z!}\lambda^{k-1}e^{-\frac{\lambda}{R}}d\lambda \\ &= \frac{1}{\Gamma(z+1)\Gamma(k)R^k}\int_0^\infty \lambda^{k+z-1}e^{-\lambda-\frac{\lambda}{R}}d\lambda \\ &= \frac{1}{\Gamma(z+1)\Gamma(k)R^k}\Gamma(z+k)\left(\frac{R}{R+1}\right)^{k+z} \\ &= \frac{\Gamma(z+k)}{\Gamma(z+1)\Gamma(k)R^k}\left(\frac{k}{R+k}\right)^k\left(\frac{R}{R+k}\right)^z \end{aligned}$$

The negative binomial offspring distribution has the following pgf in the branching process:[4,6]

$$G_z^{NB}(s) = \sum_{z=0}^{\infty}\frac{\Gamma(z+k)}{\Gamma(z)\Gamma(k+1)}\left(\frac{R}{R+k}\right)^z\left(1-\left(\frac{R}{R+k}\right)\right)^{-k}s^z = \left(1+\frac{R(1-s)}{k}\right)^{-k} \tag{2.3}$$

And the variance of the negative binomial distribution (re-parameterized for infectious disease) is:

$$Var(Z_{NB}) = R\left(1+\frac{R}{k}\right) \tag{2.4}$$

The dispersion parameter k allows us to quantify the degree of inter-individual heterogeneity and the propensity for superspreading in a population. By virtue of its

position in the variance's denominator, lower values of k ($k < 1$) correspond to higher degrees of individual heterogeneity and increasing values of k correspond to more homogeneous transmission (Figure 2.2). A negative binomial offspring distribution is particularly useful in the context of infectious disease transmission, as the epidemiologically relevant Poisson and geometric distributions are special cases of the negative binomial distribution. When $k = 1$, the variance reduces to $R(1 + R)$ and the negative binomial distribution converges to the geometric distribution. As k asymptotes to infinity, the variance reduces to R and converges to the Poisson distribution.

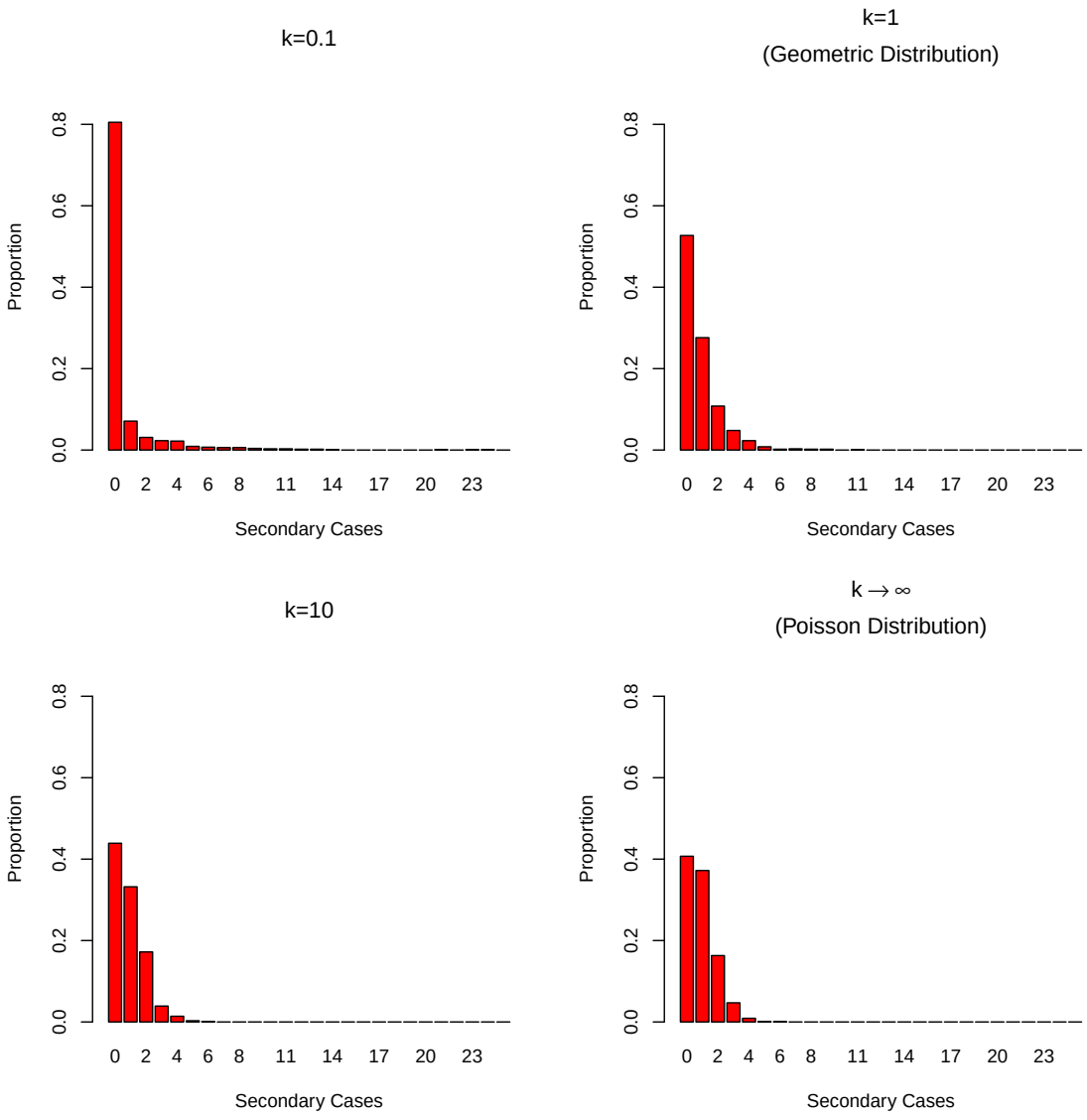

Figure 2.2: Histogram of secondary cases from 10,000 simulated transmission chains, all with identical $R = 0.9$ yet varying degrees of inter-individual heterogeneity (k). k values $<<$ 1 are result in a disproportionately high number of cases having no secondary transmission, with a minority responsible for a large number of secondary cases (high overdispersion or superspreading). As k increases above 1 and as $k \to \infty$, the distribution converges to a Poisson distribution (i.e. approaching homogeneous transmission) and differences in secondary cases are entirely attributed to stochasticity.

2.3 Parameter Inference from the Distribution of Final Transmission Cluster Sizes

In practice, generations of TB transmission are not observed with accuracy. However, total transmission cluster sizes, herein denoted Y and defined as the index case and all subsequent cases in the chain of transmission arising from the index case, are more readily identifiable. Since the chains of transmission in a surveillance system are inherently related to the final size of that chain, a natural extension of branching process theory is to base parameter inferences on the distribution of total transmission cluster sizes in a surveillance system.

While a single cluster cannot provide any insight into parameter inference, the distribution of total transmission cluster sizes in a surveillance system has been shown to be statistically sufficient in estimating the canonical parameter θ in distributions in the power series (PS) family of distributions, which includes the Poisson, geometric, and negative binomial. This holds in the context of both $R \leq 1$[11] and $R > 1$.[12] Thus, that there is no statistical efficiency gained by knowing the individual-level data (i.e. Z values) in estimating PS distributional parameters.

$$P(Z = z) = a_z \frac{\theta^z}{A(\theta)} \tag{2.5}$$

where $A(\theta) = \sum a(z)\theta^z$. Let Y represent the entire cluster size generated by the branching process, including the number of index cases, it follows that:[11]

$$P(Y = y) = \left\{ \prod_{i=1}^{y} a(z_i) \right\} \frac{\theta^{\sum z_i}}{A^y(\theta)} \tag{2.6}$$

where $i = 1, 2, 3, 4, \ldots, y$. For clusters originating with a single index cases, this is

proportional to $\frac{\theta^{y-1}}{A^y(\theta)}$, thus:

$$P(Y = y) = b(y)\frac{\theta^{y-1}}{A^y(\theta)} \tag{2.7}$$

with $b(y)$ as a proportionality constant. Importantly, a common issue with cluster level data is that occasionally transmission clusters cannot be unambiguously separated, resulting in a final cluster of size Y originating with n known index cases. Thus, the above formula can be modified such that $\frac{\theta^{y-n}}{A^y(\theta)}$ and thus:[11]

$$P(Y = y) = b(y; n)\frac{\theta^{y-n}}{A^y(\theta)} \tag{2.8}$$

This relationship has long been shown to accurately infer R using cluster level data from the single-parameter Poisson and geometric distributions.[3,11,12] These distributions are associated with inherent assumptions regarding inter-individual heterogeneity in transmission. Relatively little work has evaluated the use of final transmission cluster size distributions using the negative binomial distribution and its two parameters, R and k, which afford an unknown degree of heterogenenity.[3,12] This body of work seeks to expand on this property of branching process theory as it applies to the negative binomial distribution to more accurately infer inter-individual heterogeneity from cluster-level data.

2.3.1 Relating the Individual Offspring Distribution and the Final Cluster Distribution

As discussed, the distribution of cluster sizes is related to the distribution of secondary cases in a surveillance system. To establish a relationship between these two distributions, first recall the coefficient p_z of the pgf $G_z(s) = \sum_{z=0}^{\infty} p_z s^z$ generally specifies the probability that a single individual will infect z secondary cases. If the

probability distribution of $G_z(s)$ is given under the assumption that it is a smooth function of s with higher order derivatives, as is the case with relevant epidemiological distributions, then taking the z^{th} derivative of $G_z(s)$ and evaluating at $s = 0$ determines $P(Z = z)$; i.e. the probability that one infectious case results in z secondary infections:[3]

$$P(Z = z) = \frac{1}{\Gamma(z+1)} \frac{d^z G_z(s)}{ds^z}\bigg|_{s=0} \tag{2.9}$$

where $z = 0, 1, 2, \ldots, Y - 1$ and $\Gamma(z + 1) = z!$. To expand this concept, a common manipulation in branching processes is the multiplication of i generating functions. Thus, the coefficients of $G(s)^i$ provide all the possible ways that i cases collectively generate $0, 1, 2, \ldots, z$ cases. It follows:

$$P(Z = z|i) = \frac{1}{\Gamma(z+1)} \frac{d^z G_z(s)^i}{ds^z}\bigg|_{s=0} \tag{2.10}$$

However, extracting the probability for total cluster size, y, as a result of n index cases is not the same as the probability that z secondary infections were caused by i cases. This is because when considering all cases in a single cluster, each must be caused by another within the same chain of transmission. Therefore, for each cluster of size y with n index cases, there are y individuals that cause a total of $y - n$ secondary infections (i.e. the total cluster size minus the number of index cases). Under this constraint, only certain combinations of events generated from $G_z(s)^y$ are valid. To account for this, a normalization factor of n/y is applied to $G_y(s)$:[7,14]

$$G_y(s) = \frac{n}{y} G_z(s)^y \tag{2.11}$$

Similar to above, this normalization now allows the recovery of $P(Y = y|n)$ by dif-

ferentiating $G_y(s)$ at the $(y-n)^{th}$ derivative and evaluating at $s=0$:

$$P(Y=y|n) = \left(\frac{n}{y}\right)\frac{1}{\Gamma(y-n+1)}\frac{d^{y-n}}{ds^{y-n}}\bigg|_{s=0} \tag{2.12}$$

A closed form of $G_y(s)$ does not always exist, though may be solved recursively by the helpful conventions that $P(Y=0)=0$ as there must be at least one case in the cluster, and that $Y=1$ represents a single index case transmitting zero secondary cases, thus $P(Y=1)=P(Z=0)$:

$$P(Y=1) = \frac{d}{ds}G_y(0) = G_z(0) = P(Z=0) \tag{2.13}$$

To demonstrate the recursive procedure and the continued relationship between the generating functions of the total cluster size and individual secondary cases, consider a simple cluster generated from one index case. If the index case results in no secondary cases, $Y=1$ and $P(Y=1)=P(Z=0)$. When the index case results in a single secondary case, $Y=2$, the only permutation available is that the index case transmitted to one other person who subsequently did not transmit to anyone:

$$P(Y=2) = \frac{1}{2}\frac{d^2}{ds^2}G_y(0) = P(Z=1)P(Z=0) \tag{2.14}$$

When $Y=3$, two transmission patters are valid. The index case may cause two secondary cases and neither subsequently transmit, or the index case results in one case who subsequently results in another case that does not transmit, thus:

$$P(Y=3) = \frac{1}{6}\frac{d^3}{ds^3}G_y(0) \tag{2.15}$$

$$= \frac{1}{2}\frac{d^2}{ds^2}G_y(0)[G_y(0)]^2 + [G'_y(0)]^2 G_y(0) \tag{2.16}$$

$$= P(Z=2)P(Z=0)^2 + P(Z=1)^2 P(Z=0) \tag{2.17}$$

This relationship continues for any value of $Y \in [1, \infty]$.

This generalized relationship is useful when $G_y(s)$ takes a mathematically tangible form. Assuming a negative binomial distribution, $G_y^{NB}(s) = (1 + (R(1-s))/k)^{-k}$ the recursive procedure can regenerate the entire distribution of Z for any Y. When $Y = 1$:

$$P(Y = 1) = \frac{d}{ds} G_y(0) = \left(\frac{k+R}{k}\right)^{-k} \tag{2.18}$$

For $Y \geq 2$, the recursive calculation yields:

$$G_y^{NB(i)}(s) = \frac{\prod_{j=0}^{j-n-1}(ky+j)}{y-n}\left(\frac{R}{k}\right)^{j}\left(1+\frac{R}{k}\right)^{-ky-i+n} \tag{2.19}$$

Substituting into the $P(y = y|n)$ formula above:

$$P(Y = y|n) = \binom{n}{y}\frac{\prod_{j=0}^{y-n-1}(ky+j)}{\Gamma(y-n+1)}\left(\frac{R}{k}\right)^{y-n}\left(1+\frac{R}{k}\right)^{ky+y-n} \tag{2.20}$$

Rewriting this equation utilizing the gamma function yields the final probability distribution for the final size of a transmission cluster of size Y, with underlying Z distributed $NB(R, k)$ is given by:

$$P(Y = y|n) = \binom{n}{y}\frac{\Gamma(ky+y-n)}{\Gamma(ky)\Gamma(y-n+1)}\frac{\left(\frac{R}{k}\right)^{y-n}}{\left(1+\frac{R}{k}\right)^{ky+y-n}} \tag{2.21}$$

Importantly, the values of R and k are preserved throughout the transformation from individual to cluster generating functions, thus this density function provides the foundation for interpreting inter-individual level heterogeneity using cluster level data. This equation was computationally verified using stochastic simulation (see Appendix for code).

2.4 Chapter 2 References

1. Harris TE. The Theory of Branching Processes. Berlin: Springer, 1963.

2. Pakes AG. Ch. 18. Biological applications of branching processes. Handbook of Statistics: Elsevier; 2003:693-773.

3. Yan P. Distribution Theory, Stochastic Processes, and Infectious Disease Modeling. In: Brauer F, Driessche vd, Wu J, eds. Mathematical Epidemiology. New York: Springer; 2008.

4. Taylor H, Karlin S. An Introduction to Stochastic Modeling. San Diego, California, USA: Academic Press; 1998.

5. Diekmann O, Heesterbeek JAP. Mathematical epidemiology of infectious diseases: model building, analysis and interpretation. Chichester: John Wiley and Sons; 2000.

6. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438:355-9.

7. Blumberg S, Lloyd-Smith JO. Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. PLoS Comput Biol 2013;9:e1002993.

8. Lloyd-Smith JO. Maximum Likelihood Estimation of the Negative Binomial Dispersion Parameter for Highly Overdispersed Data, with Applications to Infectious Diseases. PLOS ONE 2007;2:e180.

9. Nishiura H, Yan P, Sleeman CK, Mode CJ. Estimating the transmission potential of supercritical processes based on the final size distribution of minor outbreaks. J Theor Biol 2012;294:48-55.

10. Greenwood M, Yule GU. An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. Journal of the Royal Statistical Society 1920;83:255-79.

11. Becker N. On parametric estimation for mortal branching processes. Biometrika 1974;61:393-9.

12. Farrington CP, Kanaan MN, Gay NJ. Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics 2003;4:279-95.

13. Farrington CP, Grant AD. The distribution of time to extinction in subcritical branching processes: applications to outbreaks of infectious disease. Journal of Applied Probability 1999;36:771-9.

14. Dwass M. The total progeny in a branching process and a related random walk. Journal of Applied Probability 1969;8:682-6.

Chapter 3

Evaluating a Method to Infer Inter-Individual Heterogeneity in TB transmission Using Cluster Level Data

3.1 Abstract

Quantifying inter-individual heterogeneity in infectious disease transmission, defined as differences in the number of secondary cases between individuals, is commonly used to improve our understanding of infectious disease transmission dynamics. Recent evidence suggests tuberculosis (TB) transmission is characterized by extreme inter-individual heterogeneity ("superspreading"). Unfortunately, the unique natural history of TB prevents the accurate identification of discrete person-to-person transmission events, thus traditional methods used to quantify the extent of inter-individual heterogeneity in TB cannot be applied. However, surveillance systems can often reasonably identify entire TB transmission clusters (an index case and all

subsequent cases). In this paper, we develop and evaluate a method that accurately quantifies inter-individual heterogeneity in transmission dynamics using TB cluster distributions, without requiring knowledge of individual-level transmission events. We further validate the robustness of the inference procedure despite limitations affecting cluster-size data, such as missing cases, contact tracing, censorship, and overlapping transmission clusters. Lastly, we demonstrate the epidemiologic utility of this method by applying it to United States surveillance data obtained from the U.S. Centers for Disease Control and Prevention, concluding that TB transmission in the U.S. is characterized by a high propensity for superspreading.

3.2 Background

With more than 10 million new cases and 1.5 million deaths in 2019, tuberculosis (TB) is a major contributor of global morbidity and mortality.[1] Despite the global TB incidence rate declining over the past twenty years, the rate of decline has recently decelerated and is now insufficient to achieve global TB targets by the end of this century.[1,2] Multiple analyses suggest that this rate of decline will continue to slow in the absence of additional and more targeted interventions.[3–5]

Incident cases of TB arise either through reactivation of a latent TB infection (LTBI) acquired in the distant past, or recent transmission. Recent transmission is distinguished from reactivation of LTBI as it focuses on the proportion of infected individuals who progress to active TB within a finite timeframe after infection (0-3 years). While recent transmission accounts for a minority of incident cases, it represents an increased potential for sporadic outbreaks that can fuel larger epidemics and lead to secondary outbreaks in other populations.[6–9] Hence, in addition to LTBI interventions, preventing recent transmission remains a key pillar in TB control programs seeking to reduce incidence.

Growing evidence suggests recent transmission is predominantly a result of "superspreading," a phenomenon wherein a small minority of infectious individuals account for the majority of secondary cases.[10,11] Such inter-individual heterogeneity in secondary cases greatly undermines interventions and is an important consideration in epidemic modeling.[12–16] Unfortunately, identifying exactly who infected whom among active cases is notoriously challenging due to the marked variability in timing from infection to clinical disease. Hence, major gaps in our understanding of inter-individual heterogeneity and its importance in shaping TB epidemiology remain.[10,11,17,18]

The inter-individual heterogeneity in a population is commonly quantified for many infectious diseases by evaluating overdispersion in the distribution of secondary cases for each infectious case.[12,19] However, these methods are not widely applicable

to TB since discrete secondary transmission events are unobserved. Fortunately, recent advances in genetic and epidemiological techniques have afforded the ability for surveillance systems to accurately identify entire TB transmission clusters (i.e. all cases in a given transmission chain).

Importantly, since individual chains of transmission give rise to the final transmission cluster size, there is a relationship between the distribution of secondary cases and the distribution of cluster sizes in a given surveillance system. Here, we evaluate a method that mathematically relates these two distributions and accurately quantifies the propensity for superspreading in TB transmission using only TB transmission cluster data. We further demonstrate the robustness of the inference procedure under complications arising in TB surveillance and demonstrate the epidemiological significance of the procedure by applying it to TB surveillance data from the United States.

3.3 Methods

3.3.1 Statistical Methods

This analysis models underlying TB transmission using a single-type branching process, also known as a Galton-Watson process. Branching processes are individual-based stochastic processes that are widely used in biology and epidemiology to study the spread of infectious diseases.[12,20–22] Analysis centers on the probability generating function (pgf) of the "offspring" distribution. The offspring distribution is the probability distribution for the number of secondary cases caused by each individual infectious case, denoted z (i.e. $P(Z = z)$ for $z = 0, 1, 2, \ldots, n$). The pgf for the number of secondary cases, z, can be generally expressed as $G_z(s) = \sum_{z=0}^{\infty} P(Z = z)s^z$.[23]

We follow previous studies in assuming that the offspring distribution follows a negative binomial distribution with mean R (the reproductive number) and dispersion

parameter k, yielding the pgf:[12,24]

$$G_z^{NB}(s) = \sum_{z=0}^{\infty} \frac{\Gamma(z+k)}{\Gamma(z)\Gamma(k+1)} \left(\frac{R}{R+k}\right)^z \left(1 - \left(\frac{R}{R+k}\right)\right)^{-k} s^z = \left(1 + \frac{R(1-s)}{k}\right)^{-k} \tag{3.1}$$

The dispersion parameter k is commonly used in epidemiology to measure the propensity of superspreading as it quantifies the degree of overdispersion in the distribution. The dispersion parameter is inversely related to the variance of the negative binomial distribution, $Var(Z_{nb}) = R(1+(R/k))$, thus smaller k values ($k << 1$) correspond to increased heterogeneity in secondary cases and imply a greater propensity for superspreading; increasing values of k correspond to more homogeneous transmission. Importantly, the negative binomial converges to the epidemiologically relevant geometric and Poisson distribution when $k = 1$ or $k \to \infty$, respectively. These distributions are often used in other applications of infectious disease transmission dynamics, such as differential equation models.

The primary focus of this analysis is to relate the offspring distribution of individual secondary cases and the offspring distribution of cluster sizes, denoted Y, in order to infer the negative binomial parameter k. This relationship is initially intuitive: consider an isolated case with no secondary transmission. The probability that a chain results in a cluster of size $Y = 1$ is identical to the probability of an individual index case having no secondary transmission, thus $P(Y = 1) = P(Z = 0)$. When expanding to a cluster of size $Y = 2$ with a single index case, the only valid transmission sequence is that an index case transmitted to a single secondary case, thus, $P(Y = 2) = P(Z = 1)P(Z = 0)$. When $Y = 3$, two valid transmission sequences can occur: either the index case transmits to two secondary cases, or the index case transmits to one secondary case, who transmits to a single tertiary case. Therefore, $P(Y = 3) = P(Z = 2)P(Z = 0)^2 + P(Z = 1)^2 P(Z = 0)$.

To expand this relationship to any cluster of size Y originating with any num-

ber of n index cases, first recall the coefficient of $P(Y = y)$ in the pgf $G_z(s) = \sum_{(z=0)}^{\infty} P(Y = y)s^z$ specifies the probability that one individual will infect z secondary cases. A common manipulation in branching process theory is the multiplication of generating functions. $G_z(s)^y$ provides all possible ways y cases can generate z secondary cases.[25,26] However, when considering transmission clusters, only a subset of these permutations result in valid transmission sequences.[27,28] To account for this, a normalization factor of n/y is applied to $G_y(s)^y$:[21,27,28]

$$G_y(s) = \frac{n}{y} G_z(s)^y \tag{3.2}$$

Bound by these constraints, the probability $P(Y = y|n)$ can be extracted by differentiating equation 2 at the $y - n^{th}$ derivative, evaluating at $s = 0$, and normalizing by $\Gamma(y - n + 1)$:[23]

$$P(Y = y|n) = \left(\frac{n}{y}\right) \frac{1}{\Gamma(y - n + 1)} \frac{d^{y-n}}{ds^{y-n}} G_x(s)^y \Big|_{s=0} \tag{3.3}$$

Assuming a negative binomial generating function (as defined in equation 3.1) yields the final probability distribution for a transmission cluster of size Y with n index cases, having underlain z values distributed $NB(R, k)$:

$$P(Y = y|n) = \left(\frac{n}{y}\right) \frac{\Gamma(ky + y - n)}{\Gamma(ky)\Gamma(y - n + 1)} \frac{\left(\frac{R}{k}\right)^{y-n}}{\left(1 + \frac{R}{k}\right)^{ky+y-n}} \tag{3.4}$$

A more detailed discussion of this relationship and programmatic code for computational validation and reproduction of this method can be found in the supplemental materials.

3.3.2 Maximum Likelihood Estimation of Transmission Parameters

Maximum likelihood estimation (MLE) was used to jointly estimate transmission parameters for both individual- and cluster-level data. Confidence intervals were obtained using profile likelihood.[29] For individual-level data, the joint likelihood is:[19]

$$L(R,k) = \prod_{z=0}^{\infty} \left[\frac{\Gamma(z+k)}{\Gamma(z+1)\Gamma(k)} \left(\frac{R}{R+k} \right) \left(1 + \frac{R}{k} \right)^{-k} \right]^{z} \tag{3.5}$$

For cluster-level data, the limitation of censoring is accounted for by designating censored clusters to be of at least size y.[30] The joint likelihood for A fully observed clusters and B censored clusters is therefore:

$$L(R,k|\vec{A},\vec{B}) = \prod_{y_a=1}^{\infty} \prod_{n_a=1}^{y_a} P(Y=y|n)^{a_{y,n}} \prod_{y_b=1}^{\infty} \prod_{n_b=1}^{y_b} P(Y \geq y|n)^{b_{y,n}} \tag{3.6}$$

where $P(Y=y|n)$ is the probability distribution function as specified in equation 3.4 and $P(Y \geq y|n) = 1 - \sum_{i=1}^{y-1} P(Y=y|n)$.

3.3.3 Simulated Data

Using this branching process framework, we simulated data to model underlying TB transmission in a surveillance system under specified values of R and k. Transmission "chains" are defined as the exact sequence of underlying transmission events (i.e. transmission trees) originating from a single index case. Transmission chains are considered to originate by the sporadic activation of latent TB or by the introduction of an infectious individual into the population (i.e. migration). A transmission "cluster" is defined as the final chain size, including the index case and all cases from all subsequent generations (i.e. secondary, tertiary, etc.) in the chain. For the purposes of this analysis, an index case with no secondary transmission is considered a "cluster"

of size 1. Each individual branching process originated with a single index case and continued until extinction. Based on empirical estimates of R for TB, our analyses focuses on values of $R < 1$, in which extinction is certain.[31,32] Specified values of $k < 1$, and particularly $k < 0.5$, are of primary interest and consistent with empirical estimates of k in TB transmission using detailed contact tracing data.[11] A simulated surveillance system consisted of N individual transmission chains. Our primary analyses simulated surveillance systems of 2000 transmission chains ($N = 2000$). Final transmission cluster sizes (Y values) were the sum of each transmission chain, including the index case. Thus, simulated cluster data were a simple vector of cluster sizes and obscured all information on individual transmission events.

3.3.4 Complications in TB Surveillance

We modeled several common real-world limitations affecting cluster size data in TB surveillance (Figure 3.1). Incomplete case ascertainment was simulated in a two-step process to emulate TB surveillance practices closely as possible. First, each case within the chain was observed with probability p_1, representing the ability of the surveillance system to passively ascertain cases (i.e. $p_1 = 1$ indicates perfect observation). Typically, once a TB case is identified in a population, many public health systems provide additional public health resources (i.e. contact tracing) to identify otherwise undiagnosed cases. Thus, to simulate active case finding all missing cases in chains with at least one case identified through passive surveillance were re-evaluated with probability p_2. After evaluation of p_1 and p_2, chains may be "broken" into two or more pseudo-clusters depending on the position of missing cases (Figure 3.1C). Censored chains were incomplete chains due to the sampling timeframe and represent ongoing transmission clusters at the time of data collection (Figure 3.1D). Each chain was designated as censored with probability p_{cens}. The generation where censoring began was randomly selected from all the generations in the chain using

a uniform distribution. The generation selected for censoring and all subsequent generations were not observed regardless of p_1 or p_2.

"Overlapping" chains considered the inability to unambiguously tease apart multiple chains and result in a combined single cluster of size y with n index cases. Overlapping chains were simulated at the cluster level by first determining the proportion of clusters in a surveillance system that overlap, p_{over}. Simulating overlap was iterative; in each iteration the process randomly drew and merged n clusters from the surveillance system, resulting in a final cluster size of $Y = \sum_n y_n$ with n index cases. The number of clusters that overlapped in each iteration (n) was drawn from a Poisson distribution with $\lambda = 1$ and then bound by a minimum of 2 ($\sim$ 70 percent) and a maximum of 7 ($<$ 0.02 percent), allowing for simulations to more accurately follow empirical estimates of the number of index cases identified from overlapping clusters.[9,33] The iterative process repeated until the proportion of chains in the surveillance system designated by p_{over} was satisfied.

Final simulated transmission chains were subject to any combination of these scenarios. "Perfect observation" was considered to be an ideal scenario where all cases in the transmission chain were perfectly observed and is the reference for the inference procedure. All simulations and calculations were completed using R statistical programming; all code needed to recreate simulations and calculations are provided in the supplemental materials.

3.3.5 United States National TB Surveillance System Data

We examined the epidemiological relevance of this method by applying the inference procedure to data from the U.S. National Tuberculosis Surveillance System (NTSS), the National Tuberculosis Genotyping Service (NTGS), and the Large Outbreaks of Tuberculosis in the United States (LOTUS) database utilized by the U.S. Centers for Disease Control and Prevention (CDC). Data are from all 50 U.S. States and the

District of Columbia, collected between January 2012 and December 2016.

Transmission cluster data were provided from the CDC using previously established methods employed by the CDC to identify likely transmission clusters described elsewhere[34]. Briefly, since 2009 the CDC has performed universal 24-locus mycobacterial interspersed repetitive unit variable number of tandem repeats (MIRU-24) in combination with obtaining clinical, demographic, geospatial, and risk factor data for all reported tuberculosis cases in the United States. Currently, the CDC uses MIRU-24 in addition to algorithms that consider risk, time, and space to identify clustered cases that may be due to recent transmission. Within this framework, the CDC further identifies large outbreaks (LOTUS; 10 or more cases in with the a 3-year period related by recent transmission) and conducts Whole Genome Sequencing (WGS) to provide increased resolution of clusters and exclude recent transmission within a given cluster.

To evaluate the sensitivity of cluster definitions, clusters were defined using two timeframes (the full 5-year data from 2012-2016 and a nested 3-year subset from 2014-2016) and two geographic scales (state and county). WGS was used to assign the number of index cases in LOTUS clusters; we did not disentangle overlapping transmission clusters within LOTUS clusters ad hoc. Clusters with an incident case arising within two years of the end of the study timeframe were considered censored.

3.4 Results

3.4.1 Initial Validation the Inference Procedure

We first compared the cluster-based inference method to the established individual-based methods[19] under perfect surveillance for 500 simulated surveillance systems, each with 2000 completely observed transmission chains (Table 3.1, Supplemental Figure S3.1). Both individual- and cluster-level data accurately inferred R and k, and

cluster-based MLE values were consistently identical or near-identical when compared to the underlying individual-level data across the range of R and k values. While perfect observation is implausible, these data verify the theoretical underpinnings of this approach and provide a basis from which the degree of bias that imperfect surveillance may impose.

3.4.2 Bias Arising Due to Complications in Surveillance

We initially evaluated the bias due to complications arising in TB surveillance univariately. Under-ascertainment of cases through passive surveillance systematically biases $\hat{k}$ upward (Figure 3.2); this was true across all values of R and k. In scenarios with extremely low passive case ascertainment ($p_1 \leq 0.3$), transmission may appear homogeneous as the distribution approximates the Poisson distribution (i.e. $k \to \infty$). Paradoxically, improving case finding through active case detection exacerbates the overestimation of k. This phenomenon is likely because the additional yield in unobserved cases from active case finding is differential with respect to cluster size; small or isolated clusters are more likely to be missed entirely by passive surveillance and subsequently not eligible for active surveillance measures (See Supplemental Figure S3.4, Panel A). Thus, improving ascertainment through contact tracing is biased towards large clusters and shifts the distribution of clusters in the surveillance system to appear more homogeneous (see supplemental materials).

Clusters that are ongoing at the time of data collection are censored. We evaluated the impact of censoring at the thresholds of 5, 10, and 20 percent of clusters censored (Figure 3.3). Censoring clusters systematically underestimated k. This is in contrast to under-ascertainment of cases due to passive and active surveillance and is likely because censoring is less likely to be differential by cluster size (see Supplemental Figure S3.4, Panel B). Our approach to addressing this limitation in the likelihood by calculating the cumulative cluster size probability of at least size Y demonstrated

modest improvement in correcting the estimates, particularly as the proportion of clusters that are censored increases.

Clusters that cannot be unambiguously isolated from other clusters are considered "overlapping," and result in a single combined cluster with multiple index cases. Inference of k is very sensitive to overlapping clusters (Figure 3.4). Without accounting for overlapping clusters in the likelihood, estimates of k were significantly biased upward (towards homogeneity). This is likely because, similar to missing cases, overlapping clusters reduces the number of isolated cases and shifts the distribution to the right. As a result, the distribution appears more normal and overdispersion is reduced. We found an *ad hoc* approach to disentangling overlapping clusters, either by evenly splitting the clusters by the number of index cases or by separating clusters such that the number of isolated cases is maximized, remained significantly biased (data not shown). However, by conditioning the likelihood on the number of index cases, we probabilistically account for all ways an overlapping cluster of size Y could be divided into n valid transmission chains. The conditional approach proved to be robust and reliably corrected for this bias across all values of k.

3.4.3 Validation of Inference Procedure Under Real World Scenarios

We evaluated the inference procedure under combined scenarios of passive surveillance, active case finding, censoring, and overlapping clusters. Based on empirical estimates of surveillance in various global settings and in consultation with TB surveillance experts, three primary scenarios were developed representing surveillance systems in high-resource, moderate-resource, and low-resource settings (Table 3.2).[35–38] Although we generate estimates across a grid of R and k values, empirical estimates are assumed to be $R = 0.5$ and $k = 0.15$.[10,11,39] Each simulated surveillance system generated 2000 transmission chains under perfect observation; after imperfect obser-

vation the median number of observed clusters in each surveillance system was 1455 (Interquartile Range (IQR): 1445-1464) for high-resource, 1151 (IQR: 1138-1163) for moderate resource, and 905 (IQR: 893-918) for low resource scenarios (data from 1000 simulations).

We found the inference of both R and k were robust and could clearly and reliably distinguished between small differences in R and k values under all scenarios (Figure 3.5). Importantly all scenarios could unequivocally distinguish between $k = 1$, which represents the geometric distribution, and all values below 0.5, including the empirical estimate of $k = 0.15$. There was a slight overestimation of k across all estimates, which systematically increases as the true underlying value of k increases. This implies the model may provide more conservative estimates of k when applied to surveillance data.

We used partial ranked correlation coefficients (PRCCs) to evaluate the strength of the relationship between each surveillance complication and its effect on k (Supplementary Figure S3.6). Under-ascertainment of cases by passive surveillance had a moderate effect and is most influential in the model (PRCC -0.594, $p < 0.001$); identifying otherwise missing cases by active case finding also had a modest effect on model estimates (PRCC -0.324, $p < 0.001$). Coverage probabilities were calculated to validate the simulation procedure for each scenario (Supplemental Table S3.1). Using the inferred R and k values from U.S. surveillance data, empirical estimates of coverage probabilities were sufficiently close to the theoretical value of 95 percent (see Supplemental Figure S3.7). Simulated confidence intervals falling outside of the true parameter were generally overestimates, implying empirical results are more conservative estimates of heterogeneity.

3.4.4 Analysis of United States TB Surveillance Data

From January 2012 to December 2016 MIRU-24 results were obtained for 95.8 percent of reported TB cases in the United States. In the full 5-year timeframe, 35,313 genotyped cases of TB were reported in the United States resulting in 29,238 clusters when defined at the county level and 26,999 clusters when defined at the state level (Table 3.3). The 3-year (January 2014 to December 2016) subset reported 20,780 cases of TB resulting in 18,128 clusters when defined at the county level and 16,212 when defined at the state level.

Inference of k remained robust throughout all four scenarios, ranging from 0.08 (3-year, state-level) to 0.12 (5-year, state-level), which is consistent with a high degree of superspreading (Table 3.4, Supplemental Figure S3.5). While $\hat{R}$ was not substantially affected by differences in the sampling timeframe, it was increased when broadening the geographic area. For the 5-year timeframe, $\hat{R}$ increased from 0.17 for county-level to 0.28 for state-level clusters and from 0.14 to 0.24 for the 3-year timeframe.

3.5 Discussion

Obtaining high resolution, individual-level data has been one of the major limitations in our understanding of TB transmission dynamics. In this study, we evaluated a method to quantify individual-level heterogeneity from more easily obtained transmission cluster data. Overall, the cluster-based inference procedure demonstrated similar accuracy in quantifying both the average transmission potential, R, and the degree of individual heterogeneity, k, when compared to individual-level data. Moreover, the inference of these transmission parameters remained robust despite real-world limitations, such as under-ascertainment of cases, overlapping clusters, and censoring of cluster size due to the study timeframe. These findings may prove useful to future surveillance efforts by drawing attention to the utility of cluster-level data

when the individual-level data unavailable.

We applied this method to TB transmission cluster data in the United States using multiple transmission cluster definitions provided by the CDC. In all scenarios, the distribution of transmission clusters was highly skewed and values of $\hat{k}$ were consistent with a high degree of superspreading. While changes in the definition of transmission clusters had a slight impact on $\hat{R}$, the effect on $\hat{k}$ was largely invariant to temporal or geospatial differences in cluster definitions (ranging from 0.07 to 0.12 between the lowest-bound and the highest-bound confidence interval). These differences are unlikely to change the epidemiological significance of the results. To our knowledge, this is the first study to quantify individual heterogeneity in TB transmission in the United States, and these results are consistent with estimates of k in other low-incidence populations outside the U.S.[10,11]

Inference of k using cluster-level data is largely a function of two key components of the distribution of clusters: the proportion of isolated cases who transmitted zero secondary cases (i.e. "clusters" of size 1) and the length of the right-hand tail. We used simulated data to investigate the direction of bias introduced from three common issues in TB surveillance and evaluated the changes in the distribution. Incomplete case ascertainment biased our estimates of k upwards. This can be explained by the fact that both passive and active surveillance methods bias case acquisition towards larger clusters; isolated cases and smaller clusters are more likely to be completely unobserved, thus shifting the distribution to the right towards homogeneity. Conversely, censoring due to the study timeframe slightly biases $\hat{k}$ downwards. This is because censoring does not affect the proportion of isolated cases and retains much of the right-hand tail, yet some clusters are indeed shifted left. Thus, the distribution appears more over-dispersed and values of $\hat{k}$ are decreased. Overlapping clusters introduce significant upward bias towards homogeneity. This is likely explained because the proportion of isolated cases is significantly reduced when combined with other

clusters, and the distribution is shifted to the right. In addition, although inference was relatively stable for all values of $k < 1$, we found that the degree of uncertainty in parameter estimation is an increasing function of k itself; larger underlying k values show broader confidence intervals around $\hat{k}$. This is likely due to the fact that, for a given R and N, as k increases individual differences in transmission become more attributed to stochasticity rather than the underlying mechanisms of transmission.

The performance of the inference procedure demonstrates that accuracy of parameter estimation is more likely a function of limitations in the data itself than by biased inference, and accurate identification of transmission clusters and index cases is paramount to the utility of these methods. This has been a historical challenge in TB surveillance using less discriminatory genotyping methods. While U.S. transmission cluster definitions were applied largely using MIRU-24 genotyping, global surveillance systems are increasingly shifting towards universal WGS of TB cases.[40] When combined with other epidemiological data, WGS provides extremely high-resolution transmission cluster data, as it can more easily differentiate between reactivation from recent transmission. Since the methods presented here are predicated on transmission clusters, the accuracy of parameter inference will increase as WGS becomes increasingly integrated in TB surveillance practices.

Our results concur with other research describing the distribution of genetic TB clusters in European surveillance data. Ypma *et al* (2013)[10] used a negative binomial branching process model to relate heterogeneity in secondary cases to the distribution of genotypic cluster sizes in the Netherlands. The authors used *IS6110* restriction fragment length polymorphism (RFLP) to define TB clusters, a much less discriminatory technique than other genotyping methods. As a result, the authors also incorporated *IS6110* transposition into their model and thus were unable to jointly estimate R and k; as the relationship between R and k is complex and nonlinear, this may affect their conclusions. More recently, Brooks-Pollock *et al* (2020)[41] investi-

gated individual heterogeneity in TB using cluster-level data assuming both negative binomial and a Poisson lognormal (PLN) distribution in describing TB cluster distributions in the Netherlands and the United Kingdom. Their PLN model showed a slightly improved fit to TB genetic cluster data but larger degree of uncertainty when compared to their negative binomial model. However, this study was looking exclusively at the distribution of TB genetic clusters (MIRU-24), not transmission clusters. Genetic clusters are typically larger than true underlying transmission clusters; further research is needed to identify distribution superiority among transmission clusters.

Our model was a simplified representation of disease transmission and subject to several limitations. Branching process models assume transmission is independent and identically distributed and we assumed heterogeneity is drawn from a negative binomial offspring distribution. Our model assumed the mean susceptibility between individuals remained constant. In reality, individual susceptibility within a population varies. In small populations with heterogeneous susceptibility, the mean susceptibility is a decreasing function of time. Highly susceptible individuals, on average, acquire infection first and thus average susceptibility reduces over generations of spread.[42] However, in both our simulated and empirical data, the population was sufficiently large such that the depletion of susceptible individuals is negligible and thus the average susceptibility does not meaningfully decline.[43] Caution should be exercised when interpreting these results in smaller populations where the depletion of susceptible individual may impact average susceptibility, which tends to decrease R.[43,44] We also assumed individual infectiousness and susceptibility were uncorrelated. Under this assumption, variation in individual susceptibility, even if unaccounted for, does not change parameter inference.[43] However, this assumption may be invalid in outbreaks occurring in vulnerable populations, such as refugee, prison, and homeless communities, where there is likely a correlation between individual infectiousness and

susceptibility. Further research should investigate the relationship between R, k and heterogeneity in individual susceptibility.

This analysis provides a well-characterized model using simplified data to infer individual differences in the number of secondary TB cases. Quantifying such information is critical to surveillance systems seeking to better understand the underlying mechanisms of TB transmission.

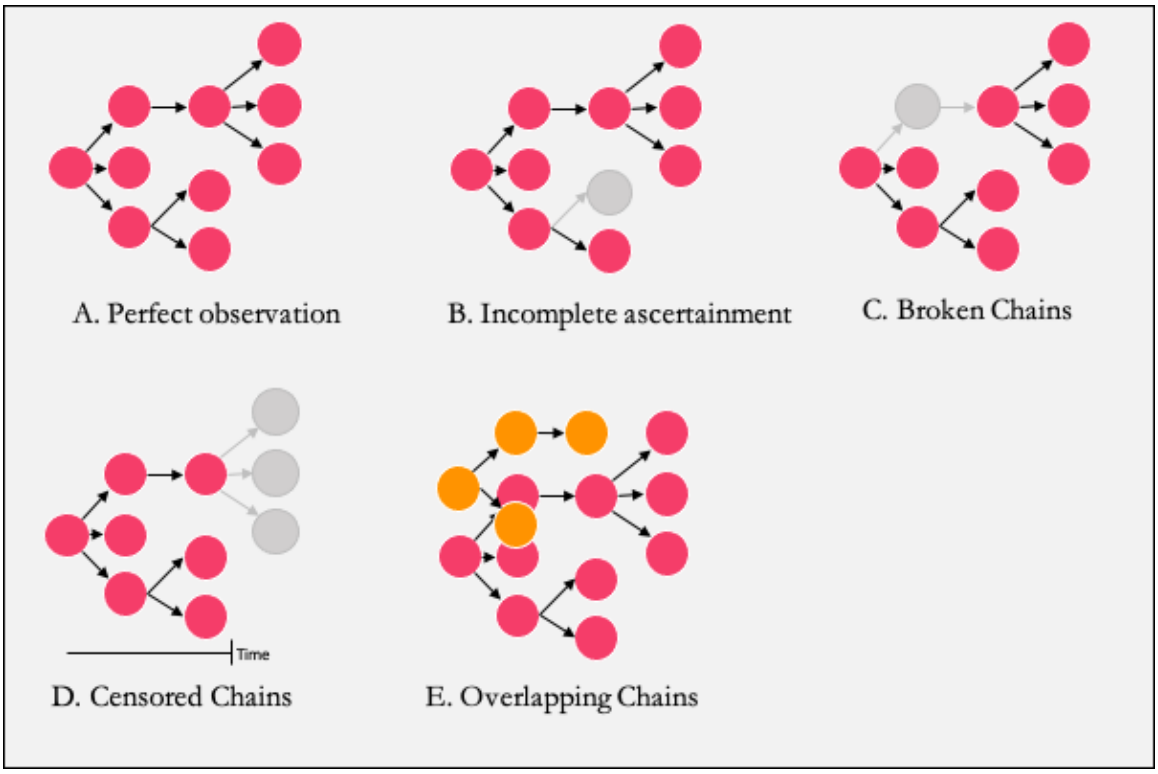

Figure 3.1: Common complications arising in TB transmission surveillance. Colored circles represent observed individuals; grey represents unobserved. Arrows represent transmission events.

Table 3.1: Individual vs cluster-level inference of k under simulated R and k values. Maximum likelihood estimates from 500 simulated surveillance systems, each with 2000 transmission chains. Coverage probabilities from simulations are in parenthesis.

	True $k = 0.25$		True $k = 0.50$		True $k = 0.75$	
	Individual	Cluster	Individual	Cluster	Individual	Cluster
True $R = 0.90$	0.25 (0.95)	0.25 (0.95)	0.50 (0.95)	0.50 (0.94)	0.75 (0.96)	0.76 (0.97)
True $R = 0.70$	0.25 (0.95)	0.25 (0.94)	0.50 (0.94)	0.50 (0.94)	0.75 (0.96)	0.75 (0.97)
True $R = 0.50$	0.25 (0.95)	0.25 (0.97)	0.50 (0.95)	0.51 (0.97)	0.76 (0.96)	0.76 (0.98)
True $R = 0.50$	0.25 (0.95)	0.25 (0.97)	0.51 (0.95)	0.52 (0.97)	0.76 (0.96)	0.78 (0.97)

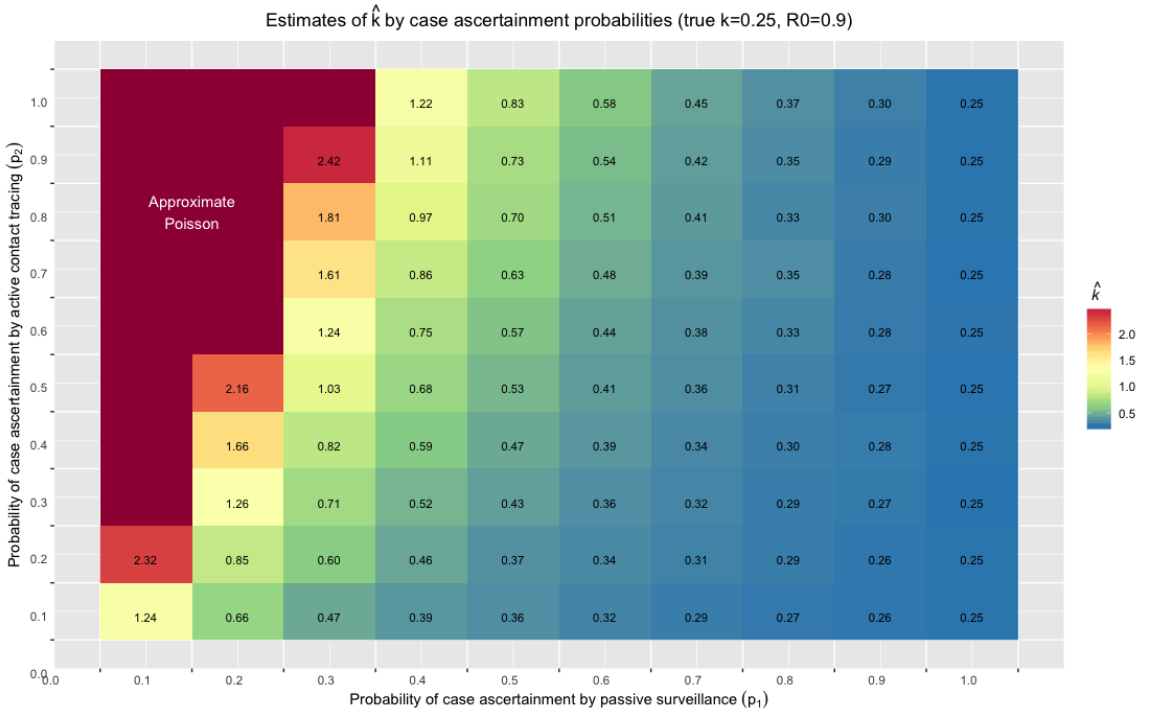

Figure 3.2: In this figure, true $k = 0.25$ and true $R = 0.90$. Passive surveillance (p_1) represents the surveillance system's ability to passively ascertain cases. Active surveillance (p_2) represents the public health system's ability to ascertain cases through contact tracing. Numbers in the center of each combination of p_1 and p_2 represent the median estimate of k of 500 simulated surveillance systems, each with 2000 chains.

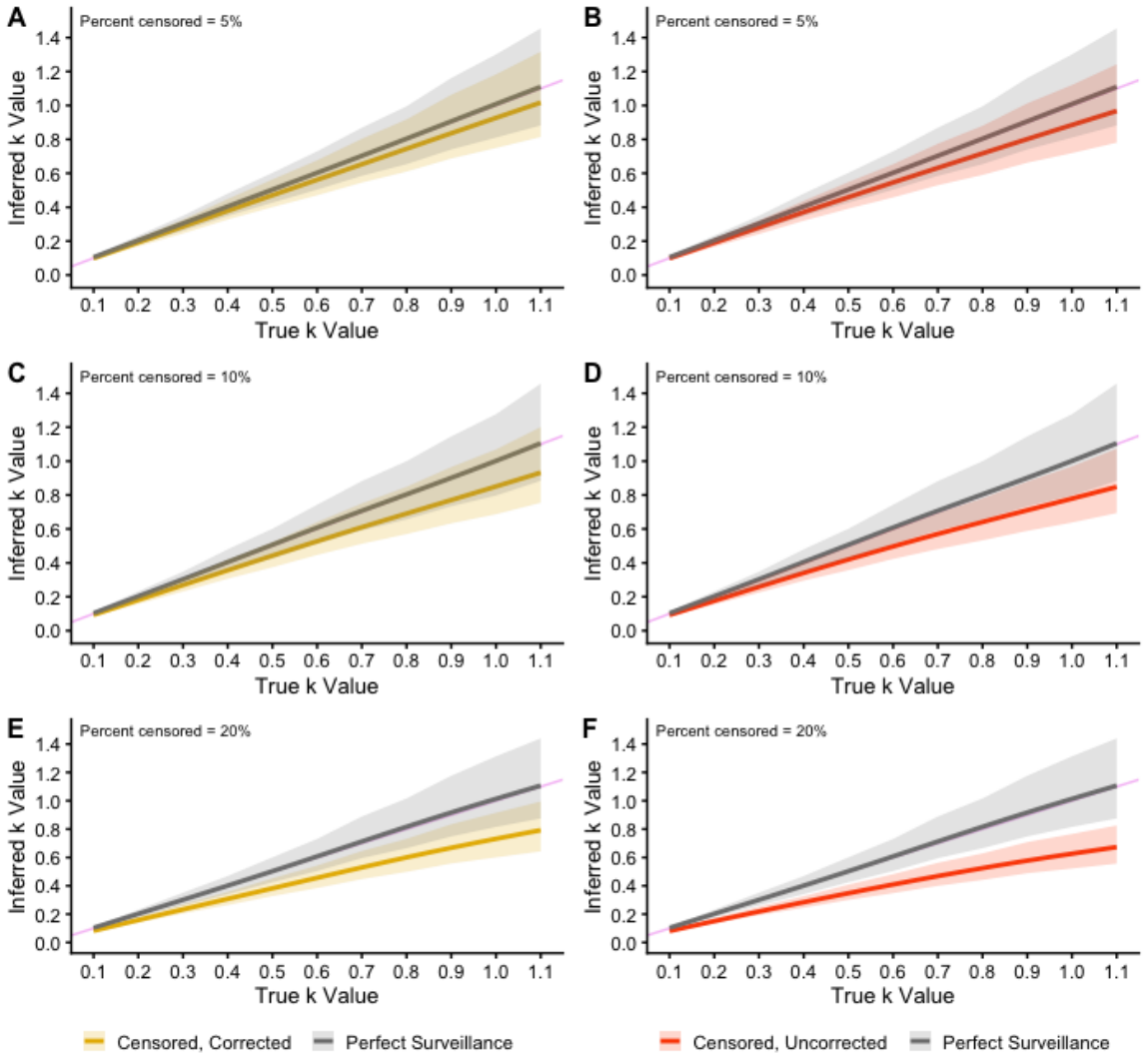

Figure 3.3: Impact of censoring on estimates of k. Results of 1000 simulated surveillance systems, each with 2000 clusters censored according to the methods, at for values of k between 0 and 1.1 ($R = 0.90$). The top row contains results when 5 percent of clusters are censored and were (A) accounted for in the final likelihood equation, or (B) unaccounted for in the likelihood. Panels (C) and (D) show similar results with 10 percent censoring, and (E) and (F) with 20 percent censoring. Grey represents the perfect observation reference (i.e. no censoring). The violet line represents perfect inference. Shading represents 95 percent confidence intervals.

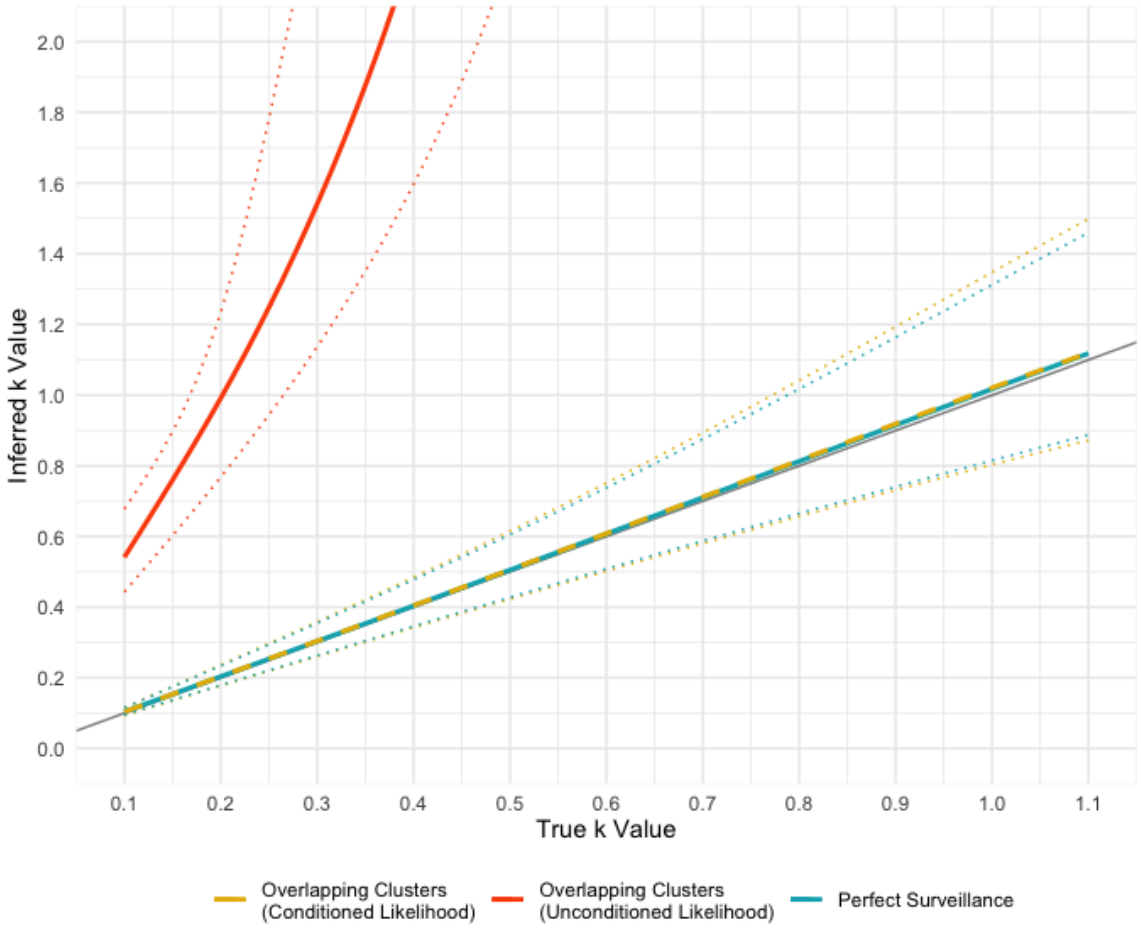

Figure 3.4: Results of 1000 simulated surveillance systems, each with 2000 clusters ($R = 0.90$). In this scenario, 20 percent of clusters overlapped ($p_{cens} = 0.20$) with the number of combined clusters ranging between 2 and 7. Dotted lines indicate 95 percent confidence intervals. The grey line indicates perfect inference.

Table 3.2: Parameter values for simulated scenarios representing high, moderate, and low resource settings.

Model Parameters	High Resource[36,37,45]	Moderate Resource[35,36,46]	Low Resource[35,37]
Proportion of cases identified via passive surveillance (p_1)	0.90	0.75	0.50
Additional yield of undiagnosed cases through active case finding efforts (p_2)	0.75	0.50	0.25
Proportion of clusters censored (p_{cens})	0.05	0.10	0.10
Proportion of clusters overlapping (i.e. with 2 or more index cases) (p_{clust})	0.15	0.20	0.20

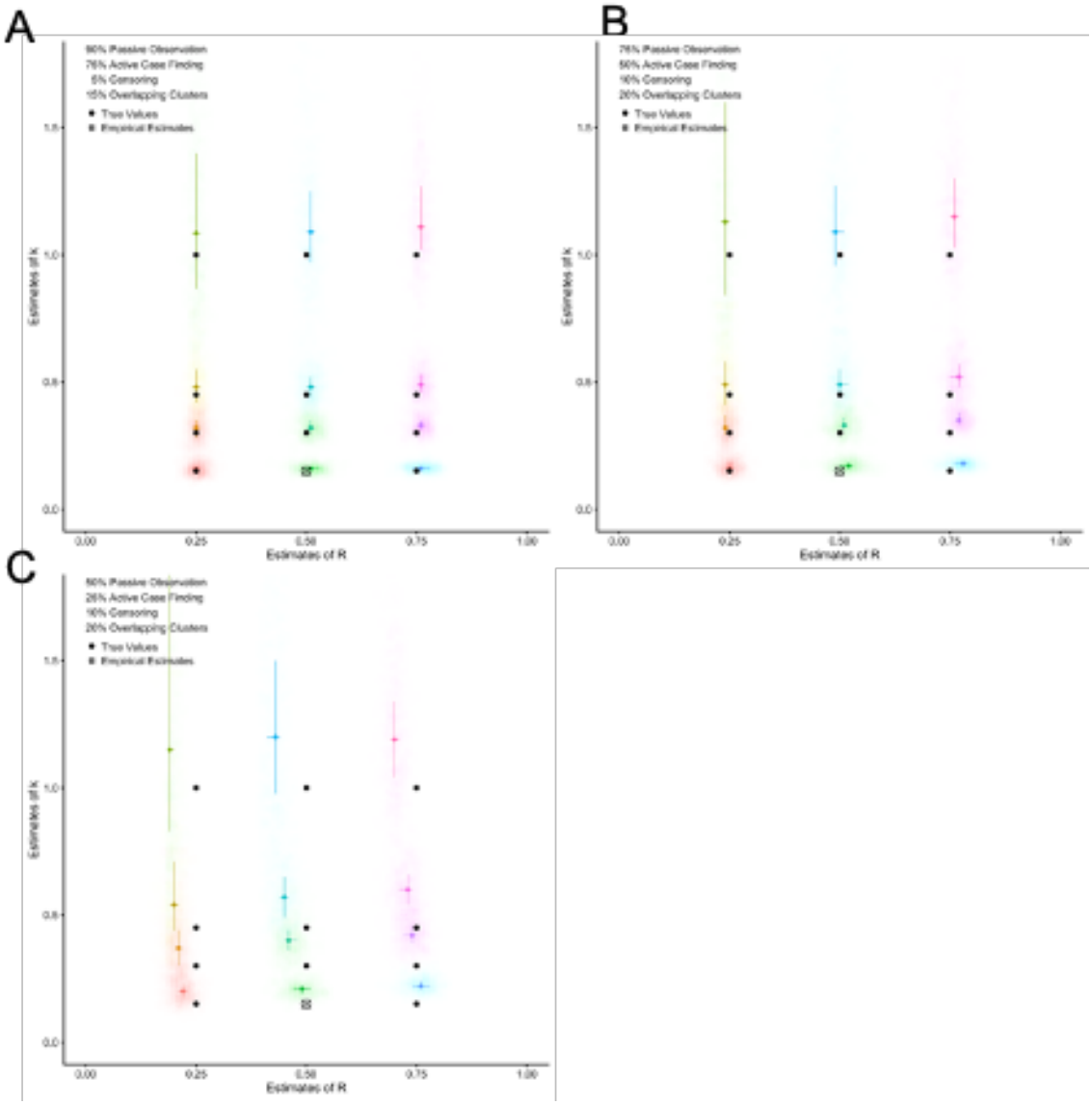

Figure 3.5: Inference from simulated surveillance systems under imperfect surveillance for A) high resource B) moderate resource, and C) low resource settings as described in the methods. Each surveillance scenario was simulated 500 times for each R and k value. Colored lines represent the interquartile range for each R and k; dots represent medians values. Black dots represent true values. R values were simulated at 0.25, 0.50 (empirical), and 0.75. k values were simulated at 0.15 (empirical), 0.30, 0.45, and 1.0.

Cluster Size	5 Years (2012 to 2016)			3 Years (2014 to 2016)		
	Number of Clusters	Total Cases	Percent of Cases	Number of Clusters	Total Cases	Percent of Cases
County-level definition						
1	26580	26580	75%	16779	16779	81%
2	1638	3276	9%	893	1786	9%
3	474	1422	4%	224	672	3%
4	203	812	2%	90	360	2%
5	98	490	1%	42	210	1%
6	66	396	1%	33	198	1%
7	52	364	1%	21	147	1%
8	29	232	1%	5	40	≤1%
9	14	126	≤1%	12	108	1%
10	12	120	≤1%	5	50	≤1%
11	12	132	≤1%	6	66	≤1%
≥ 12	60	1363	4%	18	364	2%
State-level definition						
1	22154	22154	63%	14379	14379	69%
2	1921	3842	11%	1136	2272	11%
3	629	1887	5%	291	873	4%
4	250	1000	3%	128	512	2%
5	139	695	2%	73	365	2%
6	90	540	2%	47	282	1%
7	66	462	1%	35	245	1%
8	51	408	1%	22	176	1%
9	34	306	1%	14	126	1%
10	26	260	1%	17	170	1%
11	18	198	1%	11	121	1%
≥12	1621	3561	10%	59	1259	6%

Table 3.3: Transmission cluster sizes in the United States by timeframe and geography

Sampling Timeframe	Geographic Catchment	$\hat{R}$ (95% CI)	$\hat{k}$ (95% CI)
5 Years (2012-2016)	County	0.17 (0.16-0.18)	0.09 (0.08-0.10)
	State	0.28 (0.27-0.29)	0.12 (0.11-0.12)
3 Years (2014-2016)	County	0.14 (0.14-0.15)	0.08 (0.07-0.09
	State	0.24 (0.23-0.25)	0.11 (0.10-0.12)

Table 3.4: Estimates of R and k for TB transmission in the United States by timeframe and geographic definition of clusters

3.6 Supplemental Materials

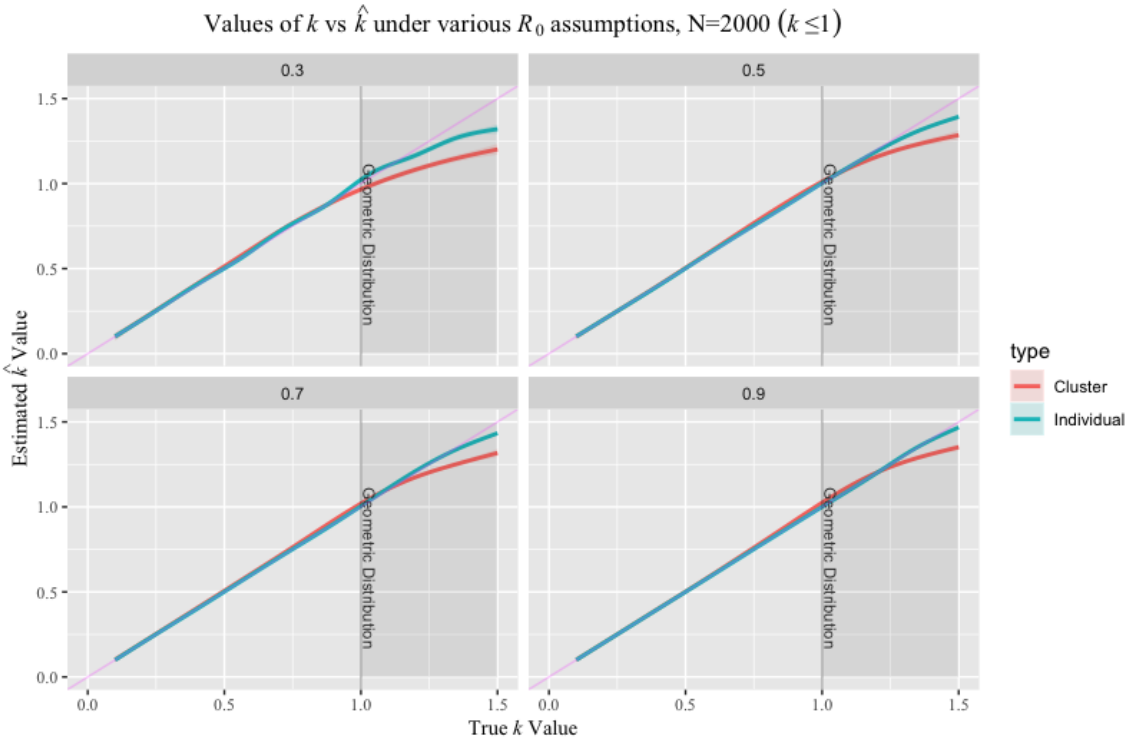

Figure S3.1: Inference of k under various R values for individual- and cluster-level data. Each surveillance system contained 2000 simulated transmission chains under perfect surveillance. Each surveillance system was simulated 100 times for underlying values of k between 0.1 and 1.5. The purple line indicates perfect inference. Values above the purple line indicate an overestimation of k; below the line indicate an underestimation of k. R values are indicated at the top of each panel. This analysis focuses on k values below 1.0; the grey shaded areas represents k values above 1.0.

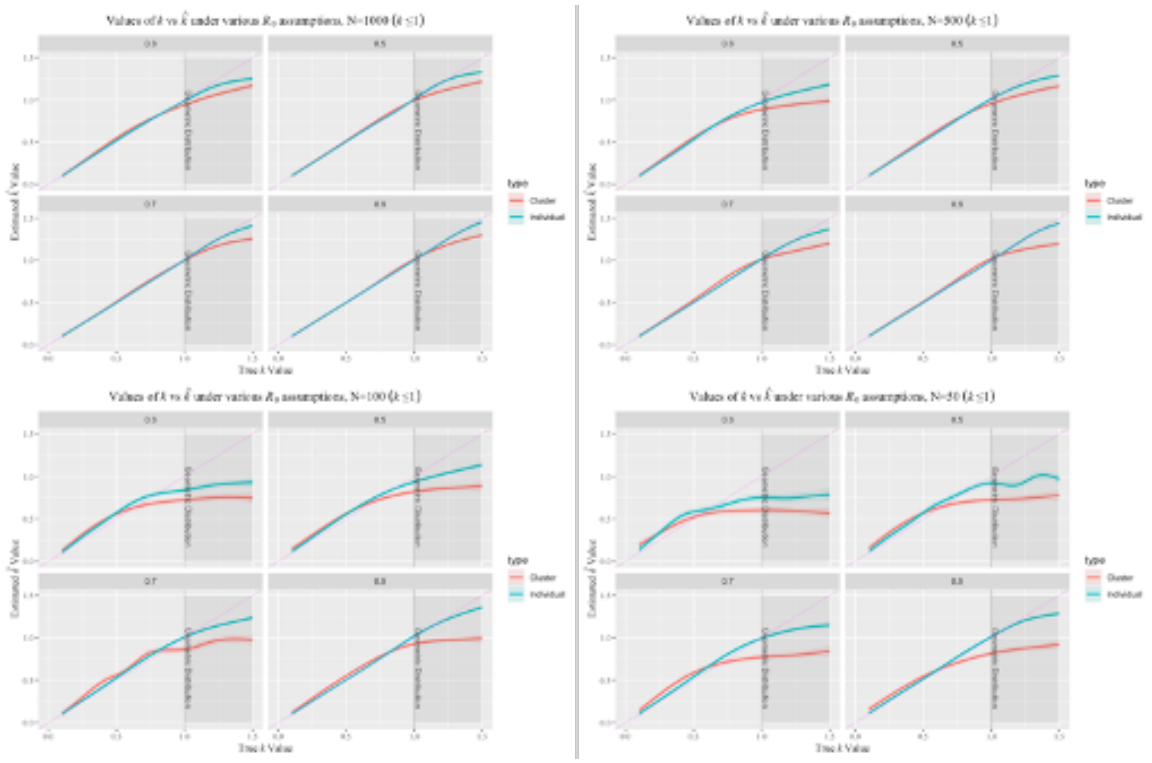

Figure S3.2: k values range from 0.1 to 1.5. Purple line indicates inferred $\hat{k}$ values are identical to the true k values. Values above the purple line indicate an overestimation of k; below the line indicate an underestimation of k. R values are indicated at the top of each panel.

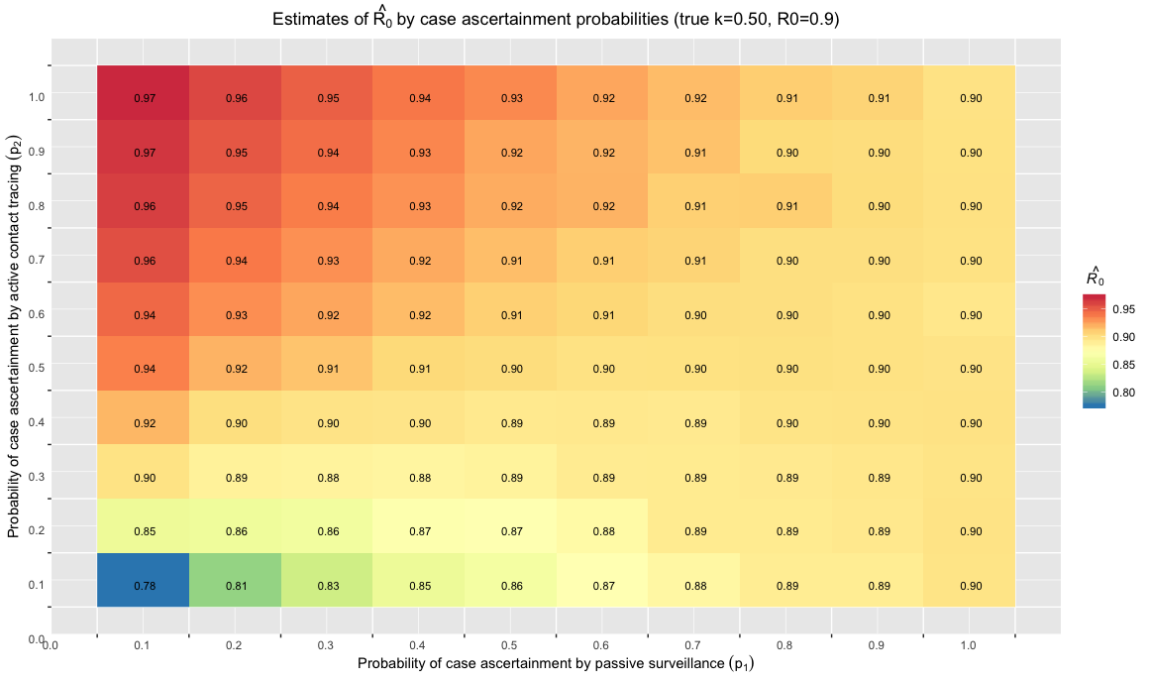

Figure S3.3: Inference of R under varying case ascertainment probabilities

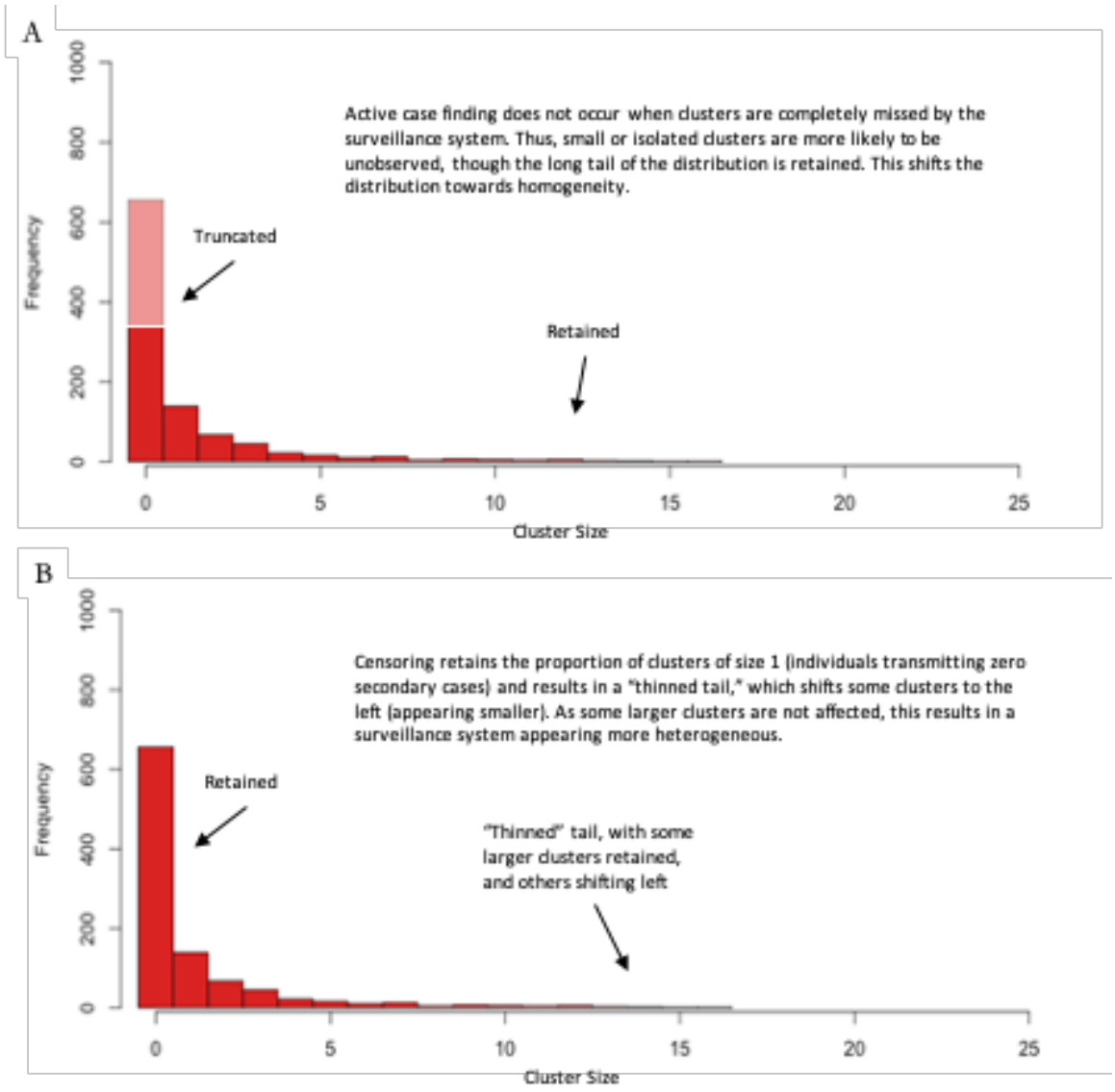

Figure S3.4: Visualizing the bias of missing cases. A) Possible explanation of bias due to imperfect case ascertainment. B) Possible explanation of bias due to censoring.

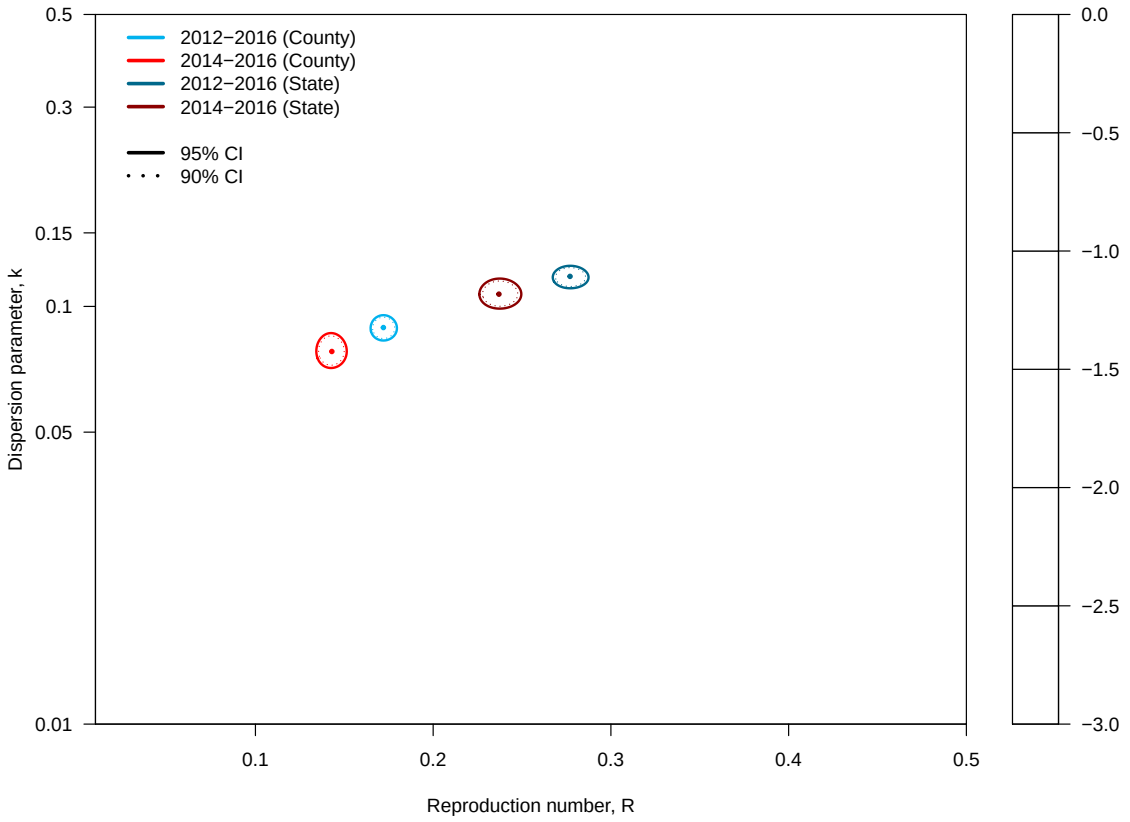

Figure S3.5: Surface Estimates of R and k in the United States

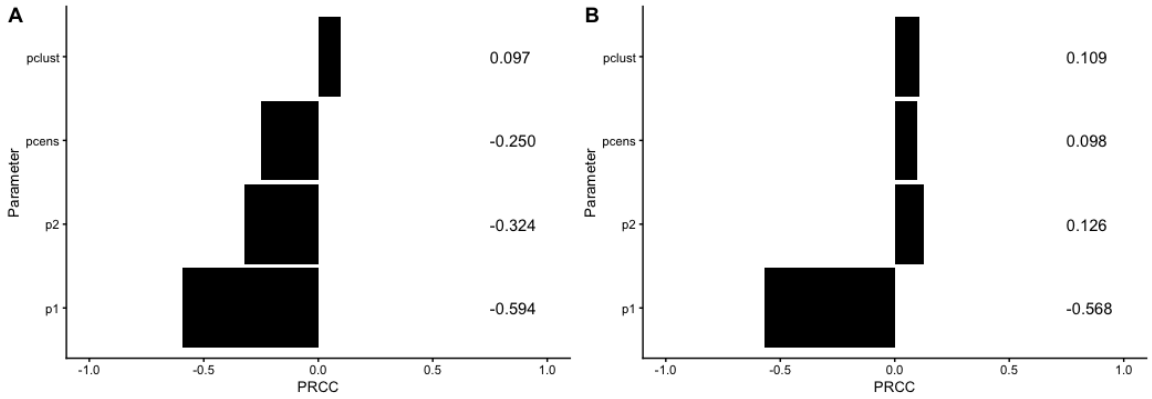

Figure S3.6: Partial rank correlation coefficients (PRCCs) comparing imperfect surveillance parameters with inference of transmission parameters. A) PRCCs for k. B) PRCCs for R.

True k value	Perfect Surveillance	High-resource Setting	Moderate-resource Setting	Low-resource Setting
0.10	0.966	0.900	0.886	0.646
0.30	0.966	0.944	0.934	0.860
0.50	0.960	0.938	0.930	0.926

Table S3.1: Coverage probabilities of the inference procedure under various k values. Each coverage probability was the result of 500 simulated surveillance systems, each with 2000 chains and $R = 0.25$.

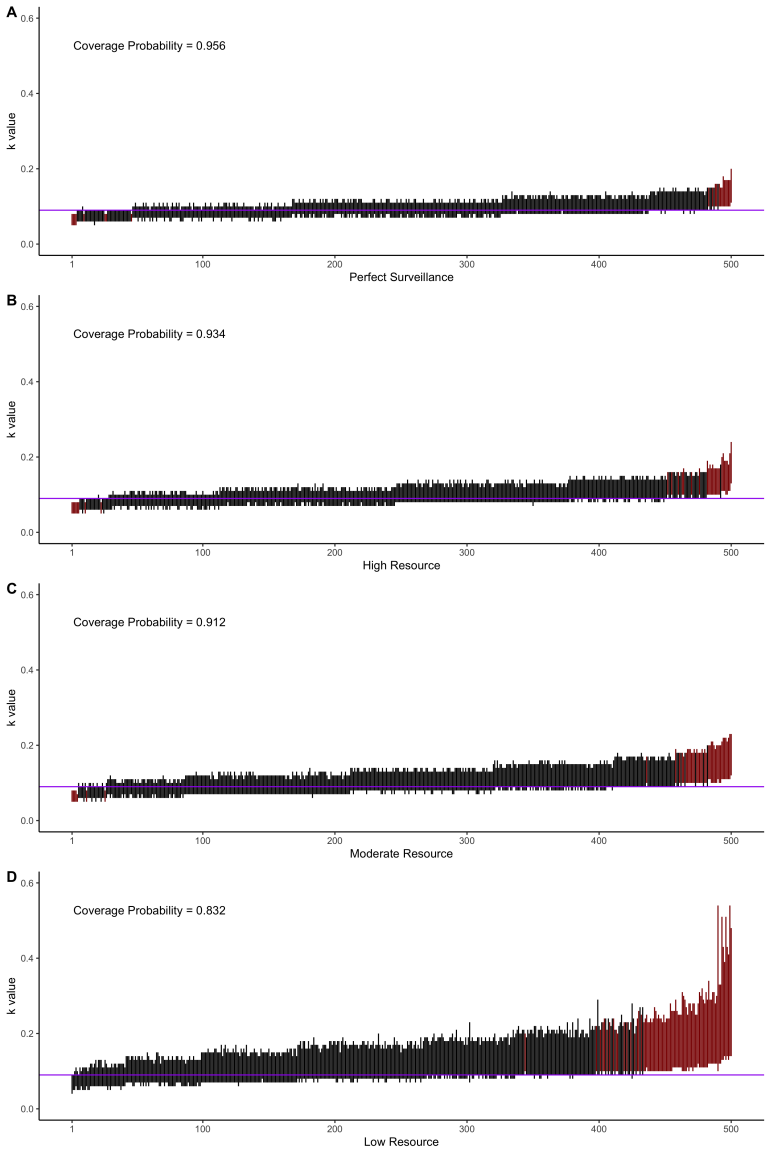

Figure S3.7: Coverage probabilities of empirical estimates of R and k using US CDC TB surveillance data, $\hat{k} = 0.09$, $\hat{R} = 0.17$. Results of 500 simulated surveillance systems, each containing 2000 transmission chains, for each scenario under MLE estimates for R and k in the United States. A) Perfect surveillance; B) High-resource setting; C) Moderate-resource setting; D) Low-resource setting. Vertical lines represent 95% confidence intervals for each simulation. The purple line represents the true parameter value of interest ($\hat{k} = 0.17$). Black represents simulations containing the true parameter in the 95% CI; red represents simulations that do not contain the true parameter in the 95% CI.

3.7 Chapter 3 References

1. Global Tuberculosis Report 2019. Geneva: World Health Organization; 2019.

2. The Stop TB Strategy. Geneva: World Health Organization; 2010.

3. Menzies NA, Cohen T, Hill AN, et al. Prospects for Tuberculosis Elimination in the United States: Results of a Transmission Dynamic Model. American journal of epidemiology 2018;187:2011-20.

4. Dowdy DW, Grant AD, Dheda K, Nardell E, Fielding K, Moore DAJ. Designing and Evaluating Interventions to Halt the Transmission of Tuberculosis. The Journal of Infectious Diseases 2017;216:S654-S61.

5. Shrestha S, Cherng S, Hill AN, et al. Impact and Effectiveness of State-Level Tuberculosis Interventions in California, Florida, New York, and Texas: A Model-Based Analysis. American journal of epidemiology 2019;188:1733-41.

6. Althomsons SP, Kammerer JS, Shang N, Navin TR. Using Routinely Reported Tuberculosis Genotyping and Surveillance Data to Predict Tuberculosis Outbreaks. PLOS ONE 2012;7:e48754.

7. Cohen T, Colijn C, Finklea B, Murray M. Exogenous re-infection and the dynamics of tuberculosis epidemics: local effects in a network model of transmission. J R Soc Interface 2007;4:523-31.

8. Mathema B, Andrews JR, Cohen T, et al. Drivers of Tuberculosis Transmission. The Journal of Infectious Diseases 2017;216:S644-S53.

9. Yuen CM, Kammerer JS, Marks K, Navin TR, France AM. Recent Transmission of Tuberculosis — United States, 2011–2014. PLOS ONE 2016;11:e0153728.

10. Ypma RJ, Altes HK, van Soolingen D, Wallinga J, van Ballegooijen WM. A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. Epidemiology (Cambridge, Mass) 2013;24:395-400.

11. Melsew YA, Gambhir M, Cheng AC, et al. The role of super-spreading events in Mycobacterium tuberculosis transmission: evidence from contact tracing. BMC

Infectious Diseases 2019;19:244.

12. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438:355-9.

13. Wong G, Liu W, Liu Y, Zhou B, Bi Y, Gao George F. MERS, SARS, and Ebola: The Role of Super-Spreaders in Infectious Disease. Cell Host and Microbe 2015;18:398-401.

14. Kucharski AJ, Althaus CL. The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. Eurosurveillance 2015;20:21167.

15. Dye C, Gay N. Epidemiology. Modeling the SARS epidemic. Science (New York, NY) 2003;300:1884-5.

16. Lipsitch M, Cohen T, Cooper B, et al. Transmission Dynamics and Control of Severe Acute Respiratory Syndrome. Science (New York, NY) 2003;300:1966-70.

17. Gardy JL, Johnston JC, Sui SJH, et al. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. New England Journal of Medicine 2011;364:730-9.

18. McCreesh N, White RG. An explanation for the low proportion of tuberculosis that results from transmission between household and known social contacts. Scientific Reports 2018;8:5382.

19. Lloyd-Smith JO. Maximum Likelihood Estimation of the Negative Binomial Dispersion Parameter for Highly Overdispersed Data, with Applications to Infectious Diseases. PLOS ONE 2007;2:e180.

20. Dubrow R. Introduction to Stochastic Processes with R. England: Wiley; 2016.

21. Becker N. On parametric estimation for mortal branching processes. Biometrika 1974;61:393-9.

22. Farrington CP, Kanaan MN, Gay NJ. Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics 2003;4:279-95.

23. Yan P. Distribution Theory, Stochastic Processes, and Infectious Disease Model-

ing. In: Brauer F, Driessche vd, Wu J, eds. Mathematical Epidemiology. New York: Springer; 2008.

24. Philippou AN. The Negative Binomial Distribution of Order k and Some of Its Properties. Biometrical Journal 1984;26:789-94.

25. Harris TE. The Theory of Branching Process. University of Michigan: Springer-Verlag; 1963.

26. Lange K. Applied Probability. Second Edition ed. New York: Springer; 2010.

27. Blumberg S, Lloyd-Smith JO. Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. PLoS Comput Biol 2013;9:e1002993.

28. Dwass M. The total progeny in a branching process and a related random walk. Journal of Applied Probability 1969;8:682-6.

29. Venzon DJ, Moolgavkar SH. A Method for Computing Profile-Likelihood-Based Confidence Intervals. Journal of the Royal Statistical Society Series C (Applied Statistics) 1988;37:87-94.

30. Farrington CP, Kanaan MN, Gay NJ. Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics 2003;4:279-95.

31. Salpeter EE, Salpeter SR. Mathematical model for the epidemiology of tuberculosis, with estimates of the reproductive number and infection-delay function. Am J Epidemiol 1998;147:398-406.

32. Borgdorff MW, Behr MA, Nagelkerke NJ, Hopewell PC, Small PM. Transmission of tuberculosis in San Francisco and its association with immigration and ethnicity. Int J Tuberc Lung Dis 2000;4:287-94.

33. Guerra-Assunção JA, Crampin AC, Houben RM, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. eLife 2015;4.

34. France AM, Grant J, Kammerer JS, Navin TR. A field-validated approach using surveillance and genotyping data to estimate tuberculosis attributable to recent

transmission in the United States. American journal of epidemiology 2015;182:799-807.

35. Saunders MJ, Tovar MA, Collier D, et al. Active and passive case-finding in tuberculosis-affected households in Peru: a 10-year prospective cohort study. The Lancet Infectious Diseases 2019;19:519-28.

36. Mor Z, Migliori GB, Althomsons SP, Loddenkemper R, Trnka L, Iademarco MF. Comparison of tuberculosis surveillance systems in low-incidence industrialised countries. European Respiratory Journal 2008;32:1616-24.

37. Doyle TJ, Glynn MK, Groseclose SL. Completeness of Notifiable Infectious Disease Reporting in the United States: An Analytical Literature Review. American journal of epidemiology 2002;155:866-74.

38. Wood R, Middelkoop K, Myer L, et al. Undiagnosed tuberculosis in a community with high HIV prevalence: implications for tuberculosis control. American journal of respiratory and critical care medicine 2007;175:87-93.

39. Ma Y, Horsburgh CR, White LF, Jenkins HE. Quantifying TB transmission: a systematic review of reproduction number and serial interval estimates for tuberculosis. Epidemiology and infection 2018;146:1478-94.

40. Cabibbe AM, Walker TM, Niemann S, Cirillo Daniela M. Whole genome sequencing of ¡em¿Mycobacterium tuberculosis¡/em¿. European Respiratory Journal 2018;52:1801163.

41. Brooks-Pollock E, Danon L, Korthals Altes H, et al. A model of tuberculosis clustering in low incidence countries reveals more transmission in the United Kingdom than the Netherlands between 2010 and 2015. PLoS Comput Biol 2020;16:e1007687-e.

42. Hougaard P. Life Table Methods for Heterogeneous Populations: Distributions Describing the Heterogeneity. Biometrika 1984;71:75-83.

43. Becker N, Marschner I. The effect of heterogeneity on the spread of disease. 1990;

Berlin, Heidelberg: Springer Berlin Heidelberg. p. 90-103.

44. Karlin S, Taylor HM. A First Course in Stochastic Processes. 2nd Edition ed. Boston: Academic Press; 1975.

45. Keramarou M, Evans MR. Completeness of infectious disease notification in the United Kingdom: A systematic review. Journal of Infection 2012;64:555-64.

46. Zhou D, Pender M, Jiang W, Mao W, Tang S. Under-reporting of TB cases and associated factors: a case study in China. BMC Public Health 2019;19:1664.

47. Haraka F, Glass TR, Sikalengo G, et al. A Bundle of Services Increased Ascertainment of Tuberculosis among HIV-Infected Individuals Enrolled in a HIV Cohort in Rural Sub-Saharan Africa. PloS one 2015;10:e0123275-e.

Chapter 4

Estimates for the Propensity of Superspreading in Tuberculosis Transmission from Global Surveillance Systems

4.1 Abstract

Background: Increasing evidence suggests recent transmission in TB is characterized by "superspreading," a phenomenon wherein a small proportion of cases account for a large number of secondary cases. Unfortunately, identifying individual transmission events in tuberculosis transmission remains elusive, thus quantifying the propensity for superspreading in a given population has been limited. However, global tuberculosis surveillance systems can reliably identify entire TB transmission clusters. The distribution of TB transmission clusters has been shown to accurately infer inter-individual heterogeneity in secondary cases in the population.

Methods: We systematically abstracted TB transmission cluster data from global surveillance systems and fit these data to a negative binomial branching process model with mechanistic adjustments to the model account for cluster size distributions. We used maximum likelihood estimation to infer the parameters R (the basic reproductive number) and dispersion k; this dispersion parameter quantifies the degree of inter-individual heterogeneity in the population.

Results: A total of five datasets were included in the study. In all studies, we found estimates of k to be consistent with a high degree of superspreading ($k << 1$). We further demonstrated that by accounting for this heterogeneity, epidemiologic models are more likely to recreate observed transmission patterns when compared to other distributional assumptions.

Conclusion: While the majority of incident TB is a result of reactivation of latent TB infection, cases that are a result of recent transmission are largely characterized by infrequent yet large outbreaks consistent with superspreading.

4.2 Introduction

Tuberculosis (TB) is the leading cause of infectious death globally.[1] While controlling TB transmission remains an essential pillar of TB control policies, relatively little is known about differences in transmission among individual cases.[2,3] For many infectious diseases, there is marked variability in the number of secondary cases caused by each individual. Such heterogeneity has been shown to greatly undermine interventions aimed at interrupting transmission and play a fundamental role in shaping epidemics.[4] A small body of evidence supports such variability in TB epidemiology, including superspreading events characterized by relatively few individuals accounting for a disproportionate number of secondary cases.[5–8] However, given the limited number of studies, two basic yet fundamental questions remain unanswered: how much heterogeneity is there among individuals in transmission, and how does this heterogeneity impact transmission dynamics?

In answering these questions, it is intuitive to focus the approach around the transmission parameter R_0 (herein referred to the more generalizable R), which describes the average number of secondary cases transmitted by an infectious individual. R has well-known fundamental and applied properties to understanding epidemic trajectory, yet as an average value cannot assess differences in the number of secondary cases between idividuals.[9] Fortunately, both R and the extent of heterogeneity can be quantified by describing the distribution of secondary cases in a given population as negative binomial with mean R and dispersion parameter k.[10] While the number of secondary cases transmitted by each individual is concentrated at the mean R, the parameter k quantifies overdispersion of the distribution. Values of $k < 1$ suggest large differences in the number of secondary cases and support the potential for superspreading, while values increasing above one indicate broadly similar transmission among individuals.

This approach is commonly used to quantify the degree of individual heterogeneity

in transmission for many infectious diseases.[4,11–13] Unfortunately, its application to TB transmission has been remarkably absent. This is primarily because a direct method of estimating k is only possible in rare instances where the transmission tree for an outbreak is known. Identifying specific individual transmission events in TB epidemiology is an enormous challenge given the large differences in time between infection and disease onset.[3,14] However, in contrast to resource-intensive individual data, TB transmission clusters are more easily identifiable and the recent expansion of whole-genome sequencing (WGS) has afforded a high degree of accuracy in discriminating between transmission clusters in a surveillance system.

In this study, we systematically gather empirical TB cluster size data from detailed contact tracing, WGS, and epidemiological surveillance of TB transmission. We jointly estimate R and k using a branching process modified to accommodate cluster size distributions to examine the extent of individual variation in secondary cases and investigate the impact such variation may have on epidemic spread. We use this dispersion parameter as a single measure from which to build a preliminary evidence base regarding the degree of heterogeneity present in TB transmission across various global contexts.

4.3 Methods

4.3.1 Search Strategy

We conducted a systematic search to identify suitable surveillance data that could be extracted from reported literature. An initial search for all English-language peer reviewed studies examining TB transmission via WGS was conducted on 23 September 2019. Medical Subject Headings (MeSH) and keywords were used to search PubMed and analogous methods were used in Web of Science, PsychINFO, and CINAHL databases. Our search strategy was informed by prior literature re-

views and expert consultation; the strategy was peer-reviewed by science librarians at the Emory Woodruff Health Sciences Center Library with expertise in systematic database searches for public health.

4.3.2 Inclusion and Exclusion Criteria

Studies were included if they satisfied the following criteria: (1) were either surveillance data or conducted empirical research using observational or experimental study designs, (2) used WGS and other epidemiological techniques, and (3) identified transmission cluster sizes and index cases. Final decisions on all included studies were confirmed by all authors. If multiple studies conducted analyses on the same (or substantially overlapping) dataset only one study was included. The decision for which study to include was made based on relevance to the aims of this study, strength of design/analysis, and confirmed by all authors. For authors who explicitly define clusters, we extracted transmission clusters and number of index cases verbatim from included studies per the author's definition. Studies that report WGS to improve contact tracing but do not explicitly define transmission clusters were included if study authors were able to elucidate the clusters from available data.

4.3.3 Parameter Inference Using Cluster-level Data

We jointly estimated R and k from cluster distributions using a maximum likelihood estimation (MLE) based method developed for TB transmission using a branching process with a negative binomial offspring distribution. The probability that n index cases result in a final cluster size of y is:

$$P(Y = y|n) = \binom{n}{y} \frac{\Gamma(ky + y - n)}{\Gamma(ky)\Gamma(y - n + 1)} \frac{\left(\frac{R}{k}\right)^{y-n}}{\left(1 + \frac{R}{k}\right)^{ky+y-n}} \tag{4.1}$$

Thus, the likelihood of R and k in a distribution of clusters having a complete

clusters of size y with n index cases, and b censored clusters of at least size y and n index cases is given by:

$$L(R,k|\vec{A},\vec{B}) = \prod_{y_a=1}^{\infty}\prod_{n_a=1}^{y_a} P(Y=y|n)^{a_{y,n}} \prod_{y_b=1}^{\infty}\prod_{n_b=1}^{y_b} P(Y\geq y|n)^{b_{y,n}} \tag{4.2}$$

Where $P(Y \geq y|n) = 1 - \sum_{i=1}^{y-1} P(Y=y|n)$. We previously validated the inference procedure and demonstrated that parameter inference of both R and k is consistent with individual-level estimates when using cluster-level data.

Importantly, the negative binomial has the beneficial property of converging to the geometric and Poisson distributions when $k = 1$ or $k \to \infty$, respectively. The geometric distribution is a distributional assumption common in epidemiologic modeling (i.e. SIR models). The Poisson distribution implies homogeneous transmission. Thus, we can determine superiority of these common distributional assumptions used in epidemiology by virtue of the confidence interval of k. Both 95% and 90% confidence intervals were obtained using profile likelihood.[15]

4.3.4 Cluster Size Probability Calculations

We calculated the expected probability that a cluster initiating with one index case would result in a final size of least size Y, i.e. $P(Y \geq y)$. For each dataset, this probability was calculated by integrating over the entire parameter surface encompassed by the study-specific R and k confidence interval. We also compared the negative binomial distribution with the geometric and Poisson distributions by integrating over the confidence interval of R and setting $k = 1$ and $k \to \infty$, respectfully. To overcome computational challenges associated with the use of infinity, k was set to 1,000,000 to approximate the Poisson. We then assessed the absolute and relative probability of observing the largest cluster in each dataset between the negative binomial, geometric, and Poisson distributions. All analyses were conducted in R statistical

software.

4.4 Results

4.4.1 Characteristics of Included Datasets

After electronic and manual search, a total of five studies met inclusion criteria and had sufficient data for extraction (Table 4.1).[16–20] Detailed information on each study is provided in the supplemental materials. Three studies[16,18,19] investigated transmission in drug susceptible TB, while one looked at multi-drug resistance (MDR-TB),[20] and one collected both MDR- and extensively drug resistant (XDR) TB.[17] All studies were from different countries; three studies were from surveillance system in low incidence settings ($\leq$50 cases per 100,000 population), including Germany,[18] United Kingdom (UK),[19] and Portugal.[17] Two studies were from higher incidence settings of China[20] and Malawi.[16]

The median timeframe for isolate collection was 6 years (range: 4-16), with a median of 247 genotyped isolates (range: 80-1687) in the surveillance system. Single nucleotide polymorphisms (SNPs), which represent a variation in a single Mtb nucleotide, are used in WGS to determine transmission clusters; the SNP threshold needed to define a transmission cluster varied by author, from 6 to 16. One study did not specify a SNP threshold.[20] A median of 37% (range: 16-66) of isolates were clustered, with a median cluster size of 3. The maximum cluster size was relatively large in two studies, with a largest cluster of 38[16] and 21[17], respectively; the remaining studies had a maximum cluster size of ¡15 (Table 4.2).

4.4.2 Transmission Parameter Estimates

Using available cluster data, we jointly estimated the reproductive number, R, and dispersion parameter, k across the surface of 90 and 95% confidence intervals (Figure

4.1). Maximum likelihood estimates of k across all datasets ranged from 0.08 to 0.34, which is consistent with a high propensity for superspreading in the population (Table 4.3). When $k = 1$, the negative binomial distribution converges to the geometric distribution, a common assumption in differential equation modeling. As $k \to \infty$, the distribution becomes Poisson which implies homogeneous transmission, thus all differences in secondary cases are attributed entirely to demographic stochasticity. All studies confidently rejected underlying mechanism of transmission consistent with homogeneous transmission. In all but one study, the 95% confidence interval for k remained below 1.0. These results imply that, in large part, recent transmission as a result of superspreading more accurately describes the mechanism of secondary transmission.

4.4.3 Cluster Size Probabilities

We calculated the probability that a single index case will result in a cluster of at least size Y, $P(Y \geq y)$ (Figure 4.2; Supplemental Figure S4.1). Independent of distributional assumption, the probability of observing larger cluster sizes naturally increases with R; however, compared to the geometric and Poisson distributions, the additional variation afforded by the dispersion parameter k considerably increased the relative probability that a large cluster would be observed under a negative binomial assumption (Table 4.4; Supplemental Figure 4.1). On average, the relative probability of observing the largest cluster was 6.6 times more likely under the negative binomial than the geometric distribution, and 27.8 times more likely when compared to the Poisson distribution. Importantly, the impact of k on the probability of the emergence of a large cluster became much more profound as estimates of R decreased and conversely attenuated as R approached one.

4.5 Discussion

The primary outcome of this study was to quantify heterogeneity in TB transmission by virtue of the negative binomial dispersion parameter k. Using available surveillance data, we found estimates of k that were consistent with a high degree of overdispersion in secondary cases. This study also sought to better understand how such variation may impact TB transmission dynamics. We found for a given R value, a smaller value of k substantially increased the probability of observing a large cluster. Taken together, these findings suggest that ongoing TB transmission may be largely fueled by the high degree of heterogeneity in secondary cases, and there exists a high potential superspreading events in TB transmission

Only two previous studies have sought to quantify individual heterogeneity using the dispersion parameter k in TB transmission. Melsew *et al*[21] used comprehensive contact investigation and long-term follow up to directly estimate k from individual-level data. While this comprehensive study provides strong support for the presence of extreme individual variation in TB transmission ($k = 0.04$), such high-resolution individual data are not feasible from a surveillance standpoint. Ypma *et al*[7] estimated k from TB cluster-level data by relating individual variation to the distribution of *IS6110* restriction fragment length polymorphism (RFLP) genotypic cluster sizes ($k = 0.10$). RFLP is less discriminatory than WGS, and methods used by the authors prevented estimation of R, which is critical to understanding the potential for superspreading given the nonlinear relationship between R, k, and cluster size (Supplemental Figure S4.1). Despite these differences, our findings concur with these studies and add to the evidence base that TB transmission may be highly over-dispersed.

Our findings were largely consistent across both high- and low-incidence settings. This indicates that, while the absolute number of cases differ, the shape of the distribution is similar in both settings. Estimates were also similar between drug sus-

ceptible and drug resistant strains. This is unsurprising, as drug resistance does not typically confer additional infectiousness.

The most important limitation of our study is the reliance on available data acquired in the published literature. Surveillance systems are likely to miss smaller clusters and isolated cases, as well as truncate ongoing clusters due to censoring of the study timeframe. Both of these complications shift the distribution toward homogeneity. Thus, bias arising from these sources is unlikely to alter our findings. We also were only able to obtain data from well-resourced surveillance systems and epidemiological studies covering large municipalities, provinces, or countries. These may not represent the true mechanics of transmission on a more local level, particularly in high-incidence settings. The number of available datasets was relatively small, and despite their concurrence, these results may not be generalizable. This study also measures the overall extent of heterogeneity in transmission and does not elucidate the mechanisms behind such variation. Superspreading remains an ill-defined term without a universally accepted definition. This study did not seek to identify superspreading events, rather quantify the potential for such phenomenon.

Two functional consequences result from these findings. First, traditional epidemiologic models accounting for this heterogeneity are more likely to reproduce the observed transmission patterns. The Poisson and geometric distributions, both common in modeling, had a significantly lower probability of recreating observed transmission patterns. Failure to properly account for such variation at an individual level may greatly undermine TB model predictions.[4] Second, we found the relationship of R, k, and total cluster size is non-linear (Supplemental Figure S4.1). As R decreased below one, the excess probability of observing ongoing transmission became more dependent on k. This implies significant potential for continued transmission despite a small R value. Thus, while R has unassailable properties in understanding transmission dynamics, estimates of inter-individual heterogeneity should be also reported

to better contextualize the mechanism of transmission in TB.

First Author (Year)	**Setting**	**Scope**	**Timeframe**	**Type of TB**
Guerra-Assunção (2015)	Karonga District, Malawi	All reported cases in location	1995-2010 (16 years)	Drug susceptible
Macedo (2019)	Portugal	All reported cases in location	2013-2017 (5 years)	M/XDR
Roetzer (2013)	Hamburg, Germany	One large 24-MIRU-VNTR cluster	1997-2010 (14 years)	Drug susceptible
Walker (2014)	Oxfordshire, UK	All reported cases in location	2007-2012 (6 years)	Drug susceptible
Yang (2017)	Shanghai, China	All reported cases in location	2009-2012 (4 years)	MDR

Table 4.1: Characteristics of Included Studies

First Author (Year)	**Culture confirmed TB cases**	**Genotyped Isolates (%)**	**SNP Cutoff**	**Unique isolates, n (%)**	**Clustered isolates, n (%)**	**Median cluster size (IQR)**	**Cluster size range**
Guerra-Assunção (2015)	2332	1687 (72)	10	672 (40)	1015 (60)	3 (2, 5)	2-38
Macedo (2019)	96	80 (83)	16	27 (34)	53 (66)	3 (2, 10)	2-21
Roetzer (2013)	86	84 (98)	3	53 (63)	31 (37)	3.5 (2.5, 5)	2-7
Walker (2014)	269	247 (92)	12	208 (84)	39 (16)	2.5 (2, 4)	2-8
Yang (2017)	367	324 (88)	Not specified	221 (68)	103 (32)	2 (2, 3)	2-8

Table 4.2: Cluster size distribution of included surveillance datasets

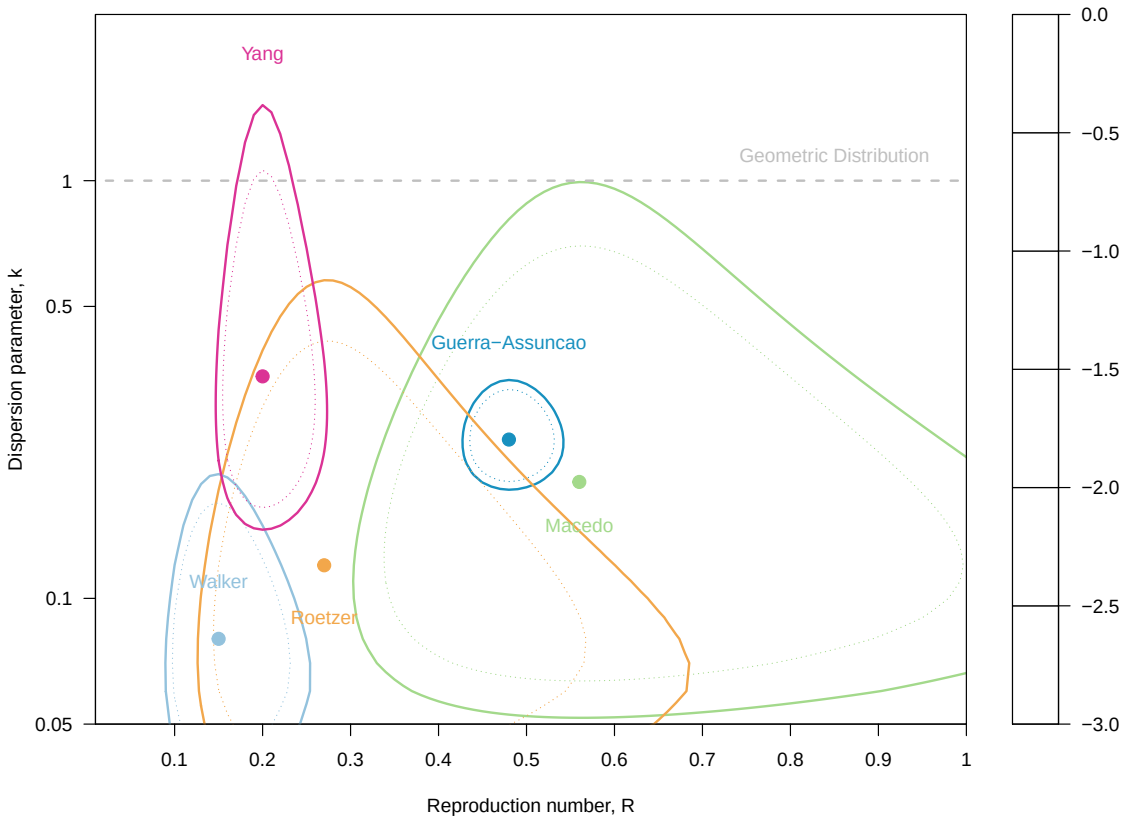

Figure 4.1: Joint estimates of the reproductive number R and dispersion parameter k for included studies. Points indicate maximum likelihood point estimates. Dotted lines indicate 90% confidence intervals and solid lines represent 95% confidence intervals.

First Author (Year)	$\hat{R}$ **(95% CI)**	$\hat{k}$ **(95% CI)**
Guerra-Assunção (2015)	0.48 (0.43-0.54)	0.24 (0.19-0.33)
Macedo (2019)	0.56 (0.31-1.05)	0.19 (0.06-0.99)
Roetzer (2013)	0.27 (0.13-0.68)	0.12 (0.04-0.57)
Walker (2014)	0.15 (0.09-0.25)	0.08 (0.04-0.19)
Yang (2017)	0.20 (0.15-0.27)	0.34 (0.15-1.51)

Table 4.3: Cluster-based maximum likelihood estimates of $\hat{R}$ and $\hat{k}$

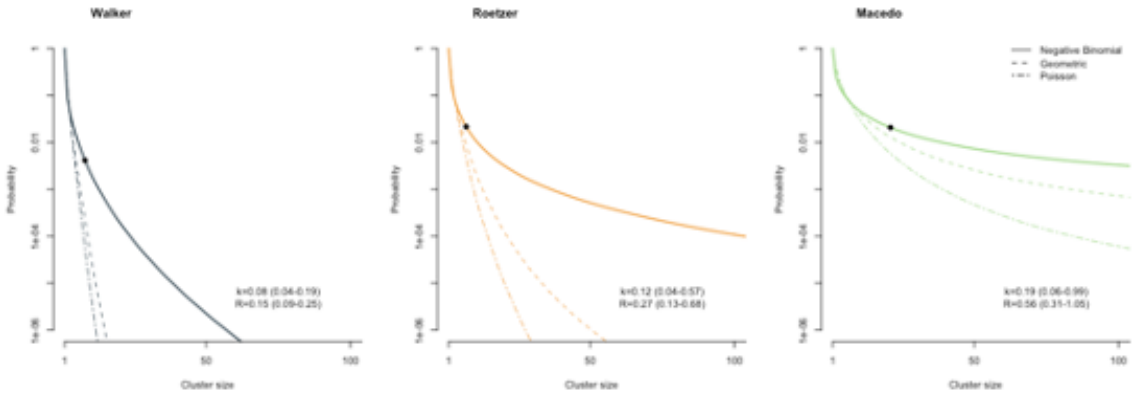

Figure 4.2: Points indicate maximum likelihood point estimates. Dotted lines indicate 90% confidence intervals and solid lines represent 95% confidence intervals.

	MLE Estimates		Absolute Probability of Observing Largest Cluster			Relative Probability of Observing Largest Cluster	
First Author (Year)	**$\hat{R}$ (95% CI)**	**$\hat{k}$ (95% CI)**	**Negative Binomial**	**Geometric**	**Poisson**	**NB vs Geometric**	**NB vs Poisson**
Guerra-Assunção (2015)	0.48 (0.43-0.54)	0.24 (0.19-0.33)	–	–	–	–	–
Macedo (2019)	0.56 (0.31-1.05)	0.19 (0.06-0.99)	0.02	0.01	0.005	1.6	3.6
Roetzer (2013)	0.27 (0.13-0.68)	0.12 (0.04-0.57)	0.02	0.008	0.004	2.4	4.8
Walker (2014)	0.15 (0.09-0.25)	0.08 (0.04-0.19)	0.004	0.0002	0.00006	19.4	73.1
Yang (2017)	0.20 (0.15-0.27)	0.34 (0.15-1.51)	0.002	0.0009	0.0003	2.1	6.3

Table 4.4: Absolute and Relative probability of observing largest cluster in observed data using alternative negative binomial, geometric, and Poisson assumptions for study data. Note the probabilities for Guerra-Assunção (2015) resulted in numeric overflow and are unable to be represented.

4.6 Supplemental Materials

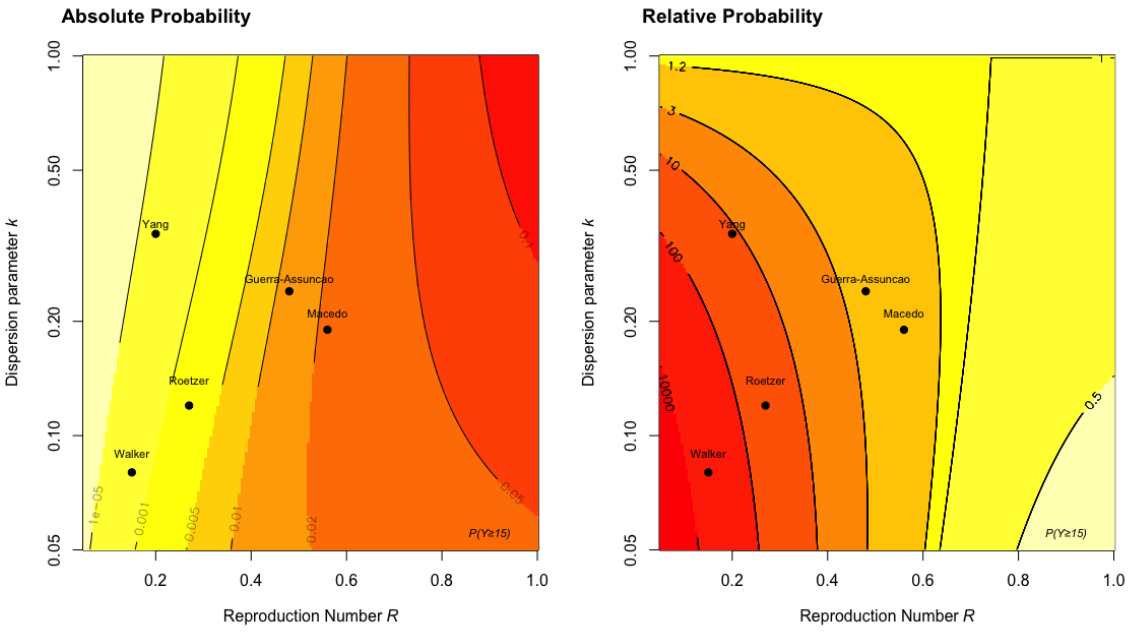

Figure S4.1: Relationship between R, k, and the probability of a cluster size of at least 15 cases, i.e. $P(Y \geq 15)$. Black dots indicate the respective author's maximum likelihood estimates of R and k. Confidence intervals are not shown for clarity. Left: The absolute probability of observing a cluster size of at least size 15 given R and k. Right: The relative probability of observing a cluster of at least size 15 compared with the geometric distribution ($k = 1$). The choice of $P(Y \geq 15)$ was arbitrary.

4.7 Chapter 4 References

1. Global Tuberculosis Report 2019. Geneva: World Health Organization; 2019.

2. Auld SC, Kasmar AG, Dowdy DW, et al. Research Roadmap for Tuberculosis Transmission Science: Where Do We Go From Here and How Will We Know When We're There? The Journal of Infectious Diseases 2017;216:S662-S8.

3. Mathema B, Andrews JR, Cohen T, et al. Drivers of Tuberculosis Transmission. The Journal of Infectious Diseases 2017;216:S644-S53.

4. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438:355-9.

5. Gardy JL, Johnston JC, Sui SJH, et al. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. New England Journal of Medicine 2011;364:730-9.

6. McCreesh N, White RG. An explanation for the low proportion of tuberculosis that results from transmission between household and known social contacts. Scientific Reports 2018;8:5382.

7. Ypma RJ, Altes HK, van Soolingen D, Wallinga J, van Ballegooijen WM. A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. Epidemiology (Cambridge, Mass) 2013;24:395-400.

8. Melsew YA, Gambhir M, Cheng AC, et al. The role of super-spreading events in Mycobacterium tuberculosis transmission: evidence from contact tracing. BMC Infectious Diseases 2019;19:244.

9. Keeling M, Rohani P. Modeling Infectious Disease in Humans and Animals: Princeton University Press; 2011.

10. Lloyd-Smith JO. Maximum Likelihood Estimation of the Negative Binomial Dispersion Parameter for Highly Overdispersed Data, with Applications to Infectious Diseases. PLOS ONE 2007;2:e180.

11. Chowell G, Abdirizak F, Lee S, et al. Transmission characteristics of MERS and

SARS in the healthcare setting: a comparative study. BMC Medicine 2015;13:210.

12. Stein RA. Super-spreaders in infectious diseases. Int J Infect Dis 2011;15:e510-3.

13. Wong G, Liu W, Liu Y, Zhou B, Bi Y, Gao George F. MERS, SARS, and Ebola: The Role of Super-Spreaders in Infectious Disease. Cell Host and Microbe 2015;18:398-401.

14. Churchyard G, Kim P, Shah NS, et al. What We Know About Tuberculosis Transmission: An Overview. J Infect Dis 2017;216:S629-s35.

15. Venzon DJ, Moolgavkar SH. A Method for Computing Profile-Likelihood-Based Confidence Intervals. Journal of the Royal Statistical Society Series C (Applied Statistics) 1988;37:87-94.

16. Guerra-Assunção JA, Crampin AC, Houben RM, et al. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. eLife 2015;4.

17. Macedo R, Pinto M, Borges V, et al. Evaluation of a gene-by-gene approach for prospective whole-genome sequencing-based surveillance of multidrug resistant Mycobacterium tuberculosis. Tuberculosis 2019;115:81-8.

18. Roetzer A, Diel R, Kohl TA, et al. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. PLOS Medicine 2013;10:e1001387.

19. Walker TM, Lalor MK, Broda A, et al. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. Lancet Respir Med 2014;2:285-92.

20. Yang C, Luo T, Shen X, et al. Transmission of multidrug-resistant Mycobacterium tuberculosis in Shanghai, China: a retrospective observational study using whole-genome sequencing and epidemiological investigation. The Lancet Infectious diseases 2017;17:275-84.

21. Melsew YA, Gambhir M, Cheng AC, et al. The role of super-spreading events

in Mycobacterium tuberculosis transmission: evidence from contact tracing. BMC infectious diseases 2019;19:244-.

Chapter 5

Estimating individual heterogeneity in tuberculosis transmission in the United States

5.1 Abstract

The decline in tuberculosis (TB) incidence in the United States (U.S.) has plateaued at a rate insufficient to achieving elimination this century. Identifying mechanisms of transmission and targeted interventions are critical to regain the historic progress made toward TB elimination. Using TB transmission cluster data from the U.S. states with the highest burden of TB, we characterized TB transmission dynamics in two ways. First, we estimated the propensity of superspreading in U.S. TB transmission by utilizing a branching process model with a negative binomial offspring distribution modified to infer individual heterogeneity using transmission cluster size distributions. The negative binomial dispersion parameter k quantifies the propensity for superspreading in the population. We then applied inferred parameters to calculate the expected proportion of cases due to recent transmission. We estimated that

the distribution of secondary cases in TB transmission was consistent with superspreading, and that the majority of incident TB is attributed to reactivation of latent TB. Major reductions in incidence can be achieved through targeted interventions dually aimed at both latent TB infection and predictively identifying and mitigating superspreading.

5.2 Introduction

Since 1989 the United States (U.S.) has pursued an ambitious goal of tuberculosis (TB) elimination, predicated largely on strategies to mitigate secondary transmission.[1,2] As a result, the U.S. has seen a marked decline in TB incidence and in 2019 reported the lowest rate in its history (2.7 cases per 100,000 population).[3] However, in recent years the rate of decline has plateaued and remains insufficient to achieving TB elimination this century.[3]

Incident cases of TB arise from either the sporadic reactivation of an infection acquired in the distant past (latent TB infection or LTBI), or recent transmission from an infectious case. While efforts to mitigate reactivation are crucial to reducing incidence, recent TB transmission is of particular concern to public health as it is holds potential for explosive outbreaks – particularly in vulnerable populations.[4,5] Such transmission patterns may be a result of "superspreading," a loosely defined term in which a minority of individual cases account for the majority of secondary transmission. Superspreading has been shown to greatly undermine prevention efforts, sustain ongoing transmission, and fuel larger epidemics.[6,7] Hence, quantifying the propensity of superspreading in a population is critical to informing intervention strategies and accelerating progress towards elimination.

Understanding the distribution of secondary cases caused by each infectious individual is useful in quantifying the possibility of superspreading in a population. For

many directly transmitted infectious diseases, this distribution follows the negative binomial distribution with mean R_0 (herein referred to the more generalizable R) and dispersion parameter k.[6–12] While the reproductive number R represents the average number of secondary cases caused by each infectious case, the dispersion parameter k quantifies the extent of variation in the number of secondary cases between individuals (i.e. specifies the degree of overdispersion in the distribution). Smaller k values ($k < 1$) indicate increased variability and thus a higher probability of superspreading. Increasingly larger values of k indicate transmission is more homogenous and less likely to be characterized by superspreading. Thus, while R remains central to our current understanding of infectious disease dynamics, k provides critical insight into transmission dynamics of infectious diseases.

In this analysis, we estimate the propensity for superspreading in TB transmission across the four U.S. states accounting for the majority of TB cases. We use a mathematical model defined by R and k of a negative binomial branching process with mechanistic adjustments to account for transmission cluster size distributions. We fit this model to the cluster size distributions of four states in the United States that comprise the majority of TB cases in the U.S. (California, Florida, New York, and Texas) to estimate the dispersion parameter k. By inference of k, we quantify the extent of individual-level heterogeneity and interpret these results in the context of superspreading. We further contextualize these findings by calculating the proportion of cases responsible for a given proportion of transmission using estimates of k inferred from the model.

5.3 Methods

5.3.1 Data Source

We used routinely collected data from the U.S. Centers for Disease Control and Prevention (CDC) National Tuberculosis Surveillance System (NTSS), the National Tuberculosis Genotyping Service (NTGS), and the Large Outbreaks of Tuberculosis in the United States (LOTUS) databases from January 1, 2014 to December 31, 2016 for the states of California, Florida, New York, and Texas. Since 2009 the CDC has performed universal 24-locus mycobacterial interspersed repetitive unit variable number of tandem repeats (MIRU-24) genotyping in combination with clinical, demographic, geospatial, and risk factor data for all reported tuberculosis cases in the U.S. Currently, the CDC uses MIRU-24 in addition to algorithms that consider risk, time, and space to identify cases that may be due to recent transmission.

5.3.2 Inference Procedure and Model

In this context, "heterogeneity" in transmission refers to differences in the number of secondary cases caused between infectious cases. This definition encapsulates the entire infectious history that led to a secondary case and may differ from other definitions of heterogeneity in infectious disease, such as those describing substantial differences in transmission between populations. Following previous studies, we assume the number of secondary cases caused by each individual is identically and independently distributed random variable according to a negative binomial distribution with mean R and dispersion parameter k.[6,13,14] The negative binomial dispersion parameter k quantifies the degree of overdispersion in the distribution and infers the propensity for superspreading in the population.

The exact number of secondary cases for each individual case is unobserved in TB transmission, however advancements in genetic techniques have allowed for surveil-

lance systems to reasonably identify entire transmission clusters, defined as the index case and all subsequent (secondary, tertiary, etc.) cases arising from the index case. Within a surveillance system, the distribution of transmission cluster sizes has been well-established as a sufficient statistic for parameter inference in power-series distributions, which contains the negative binomial.[15,16] Thus, mechanistic adjustments were made to the negative binomial probability density function (PDF) in the branching process that affords parameter inference using the distribution of final transmission cluster sizes. We made further adjustments to account for two common limitations with cluster size distributions in TB transmission. First, in some circumstances it is impossible to unambiguously separate overlapping chains of transmission. This results in a combined transmission cluster of total size Y with n index cases. In this analysis, we did not seek to disentangle overlapping transmission clusters *ad hoc*; instead accounted for this by conditioning the modified PDF on the number of index cases in the cluster. The final PDF of a transmission cluster of size Y initiating with n initial index cases can be expressed:

$$P(Y = y|n) = \binom{n}{y} \frac{\Gamma(ky + y - n)}{\Gamma(ky)\Gamma(y - n + 1)} \frac{\left(\frac{R}{k}\right)^{y-n}}{\left(1 + \frac{R}{k}\right)^{ky+y-n}} \tag{5.1}$$

Second, clusters with ongoing transmission at the time of data acquisition may result in cluster sizes being right-censored. We account for this limitation by assuming censored clusters are of at least size Y. Thus, the final likelihood for A clusters that were completely observed, and B clusters partially observed due to censoring is:

$$L(R, k|\vec{A}, \vec{B}) = \prod_{y_a=1}^{\infty} \prod_{n_a=1}^{y_a} P(Y = y|n)^{a_{y,n}} \prod_{y_b=1}^{\infty} \prod_{n_b=1}^{y_b} P(Y \geq y|n)^{b_{y,n}} \tag{5.2}$$

Where $P(Y = y|n)$ is the PDF specified in equation 3.1 and $P(Y \geq y|n) = 1 - \sum_{i=1}^{y-1} P(Y = y|n)$. We used maximum likelihood estimation (MLE) of $\hat{R}$ and $\hat{k}$ and 95% confidence intervals (CIs) were obtained using profile likelihood.

5.3.3 Transmission Cluster Definitions

The CDC provided cluster data based on CDC standard practices for identifying potential transmission clusters. Briefly, transmission clusters were defined as cases with identical MIRU-24 profiles within the same county during the study timeframe. However, the CDC further investigates LOTUS clusters (10 or more cases within a 3-year period related by recent transmission) using whole genome sequencing (WGS). WGS provides significantly higher resolution genotypic data and may identify 'overlapping" transmission clusters, where multiple transmission clusters are present in the same MIRU cluster. We also assessed the impact of varying this definition on parameter inference by expanding geographic catchment to the state level and repeating the analysis. This expanded definition intentionally provides a more conservative estimate of k.

5.3.4 Burden of Secondary Transmission

We took $\hat{R}$ and $\hat{k}$ to to specify the exact PDF and cumulative density function for the given populations. We then calculated the expected proportion of transmission attributed to a specified proportion of the cases, p_t. A general example for sexually transmitted and vector-borne diseases is the "80/20 rule" wherein 80% of transmission is due to only 20% of infectious cases (thus $p_t = 0.8$).[17] For any value of p_t, this proportion can be specified as:

$$1 - p_t = \frac{1}{R} \int_0^x u f(u) du \tag{5.3}$$

Where $f(u)$ represents the PDF of the negative binomial distribution with specified R and k per our model. In this analysis, we calculated the proportion of cases responsible for 80%, 85%, 90%, 95%, and 100% of secondary transmission. We further compared TB transmission to the common "80/20 rule" in more detail by using the MLE of k

and varying R across all values from 0 to 1 (in 0.01 increments). These calculations were computationally eased by the following manipulation:

$$1 - p_t = \frac{1}{R}\int_0^x uf(u)du = \int_0^{x-1} f(u)du \tag{5.4}$$

All analyses were performed in the R programming language; reproducible code for all analyses available in the supplementary materials.

5.4 Results

Between 2014 and 2016, a total of 21,110 incident cases of TB were reported and genotyped in the U.S. Among these, 10,970 (52%) were in the four states of California, Florida, New York, and Texas (Table 1). The vast majority of cases were isolated cases (69-83%); among clustered cases, the median cluster size was 2 (interquartile range: 2-3) for all states. The distribution of cluster sizes was heavily skewed in each state, with a vast majority of cases resulting in relatively little to no secondary cases, yet several substantially large clusters were present (Figure 5.1; Supplemental Table S5.1). This overdispersion was reflected in the MLE estimates of k; across all populations, $\hat{k}$ was substantially lower than 1 and consistent across states, ranging slightly from 0.08 to 0.11 (Table 5.2, Figure 5.2). Values of $\hat{k}$ in this range ($k << 1$) indicate extreme heterogeneity in the number of secondary cases resulting from each infectious individual and are consistent with a high probability of superspreading. Estimates are R were not the primary aim of this analysis but demonstrated slightly more variability (range: 0.14-0.22; Table 5.2, Figure 5.2). When defining transmission clusters at the state level, estimates of k were slightly higher ($\hat{k} = 0.10 - 0.17$; see Supplemental Table S5.2 and Supplemental Figure S5.1). Given the sufficiently low values of $\hat{k}$, these differences are not epidemiologically relevant as they infer a similar propensity for superspreading in the population.

We used $\hat{R}$ and $\hat{k}$ to calculate the proportion of cases responsible for 80%, 85%, 90%, 95%, and 100% of secondary transmission (Table 5.3). Our model suggests that across all four states, all secondary TB transmission ($p_t = 1.0$) is largely driven by a small fraction of infectious individuals; across all four states, only 9%-11% of infectious cases account for 100% of secondary transmission. Our model also suggests that TB transmission is more extreme than the common "80/20 rule," as across all states only 5-7% of cases are responsible for 80% of transmission. This phenomenon holds when varying R between 0-1; the number of cases responsible for 80% of infection remains under 10% in all four states for all values of R. Furthermore, these proportions did not meaningfully change when expanding the cluster definition to the state level (Supplementary Table S5.3 and Supplemental Figure S5.2).

5.5 Discussion

In this analysis, we sought to describe and quantify the degree of heterogeneity in TB transmission across the four most populous states in the United States using the negative binomial dispersion parameter k. We found transmission in all four states was characterized by a similarly high degree of superspreading, and the majority of secondary transmission is likely caused by a small minority of cases. While to our knowledge k has not been estimated for TB transmission in the United States, these results concur with estimates of k in other studies among low-incidence populations. Previous studies in Australia and the Netherlands found estimates of k of 0.04 (95% CI:0.03–0.05) and 0.10 (0.09-0.12), respectively.[13,14] Moreover, a recent study modeling TB transmission under both a negative binomial and Poisson-lognormal assumption estimated all secondary TB cases were caused by only 16% of cases in the United Kingdom, and 12% of cases in the Netherlands.

The epidemiological significance of these findings is three-fold. First, both the high

degree of overdispersion and small proportion of cases responsible for total transmission reinforces the need to focus interventions on efforts that mitigate superspreading. As the overwhelming majority of infected individuals do not lead to additional cases, incidence of TB attributable to recent transmission could be disproportionately impacted by preventing relatively rare superspreaders. Second, this analysis demonstrates the utility of cluster-level surveillance data in quantifying individual-level heterogeneity without the need for resource intensive individual data. Lastly, our findings have implications for improved TB transmission modeling. While heterogeneity in mathematical modeling is historically assigned according to some known property (i.e. smear status, HIV status), superspreading is unpredictable and often cannot be identified using *a priori* information. A notable example is a 9-year old child who was assumed to transmit TB to at least 56 contacts, while his twin brother had a relatively mild case and was not considered infectious.[18] Several models have begun incorporating stochasticity of individual variation in secondary transmission by assigning a random variable drawn from a distribution with mean R and dispersion k.[19–21] By establishing empirical estimates of the dispersion parameter in U.S. TB transmission, individual heterogeneity may be more accurately incorporated into future mathematical models that seek to further characterize TB transmission dynamics and prevention measures in the United States.

Our analysis was subject to several notable limitations. States may differ in their surveillance and intervention capacities, which invariably modulates estimates of overdispersion. However, given the marked similarity in $\hat{k}$ across all four states, any bias arising due to such differences may be insignificant as it pertains to this analysis. However, while the degree of superspreading did not vary meaningfully across states, the clinical and structural determinants of such heterogeneity may vary throughout populations. Local surveillance systems should identify characteristics that play a key role in superspreading for their specific populations. While we accounted for

truncated cluster size due to censoring in the likelihood, we did not account for systematic under reporting of cases (or cases missing genotype). In the context of cluster-based inference, such missing cases may bias inference towards homogeneity (increased $\hat{k}$),thus our estimates are likely conservative. This bias arises because parameter inference is largely predicated on two key components of the distribution: the long right-hand tail of the distribution and the proportion of isolated cases. When a case of TB is identified in the United States, this often triggers additional public health resources for active case finding (i.e. contact tracing). By doing so, this intuitively biases case ascertainment differentially towards larger clusters. Larger clusters are exponentially more likely to have at least one case identified, thereby triggering active case finding efforts and capturing otherwise unidentified cases. In contrast, smaller clusters – particularly orphan cases with no secondary transmission – are more likely to be missed entirely by the surveillance system. This phenomenon shifts the distribution to the right towards homogeneity.

Although disease transmission is not limited to administrative borders, our primary analysis imperfectly defined transmission cluster using county-level data within each state. While this definition provides the most reasonable identification of transmission clusters give data availability, it likely misclassifies larger transmission clusters as multiple smaller clusters. Such misclassification biases $\hat{k}$ downwards towards heterogeneity, as it increases the proportion of cases that appear to transmit zero secondary cases. To address this, we expanded our definition to the state level, which provides a very conservative estimate of cluster sizes. We found that the inference of k was largely invariant to the change in definition, with only marginal increases in $\hat{k}$ in each state. Thus, the epidemiological relevance of our findings likely does not change based on transmission cluster definition.

Understanding the degree of individual variation in the number of secondary transmissions is crucial for epidemic control. Our findings suggest that there is considerable

variation in the capacity for individuals to transmit TB in the United States, and most cases do not contribute to ongoing transmission. In addition to interventions aimed at reducing TB reactivation, efforts aimed at preventing superspreading will likely meaningfully contribute to the goal of elimination of TB in the United States.

Jurisdiction	**Total Cases**	**Proportion of all reported cases in the U.S.**	**Proportion of Isolated Cases**	**Proportion of Clustered Cases**	**Median Cluster Size (IQR)**	**Largest Cluster Size**
California	5,024	24%	0.75	0.25	2 (2, 3)	41
Florida	1,415	7%	0.79	0.21	2 (2, 3)	42
New York	1,757	8%	0.83	0.17	2 (2, 3)	14
Texas	2,774	13%	0.69	0.31	2 (2, 3)	65
Total (All Four States)	10,970	52%	0.75	0.25	2 (2, 3)	65
Entire United States	21,110	100%	0.79	0.21	2 (2, 3)	65

Table 5.1: Reported TB cases and clusters in the United States, 2014-2016

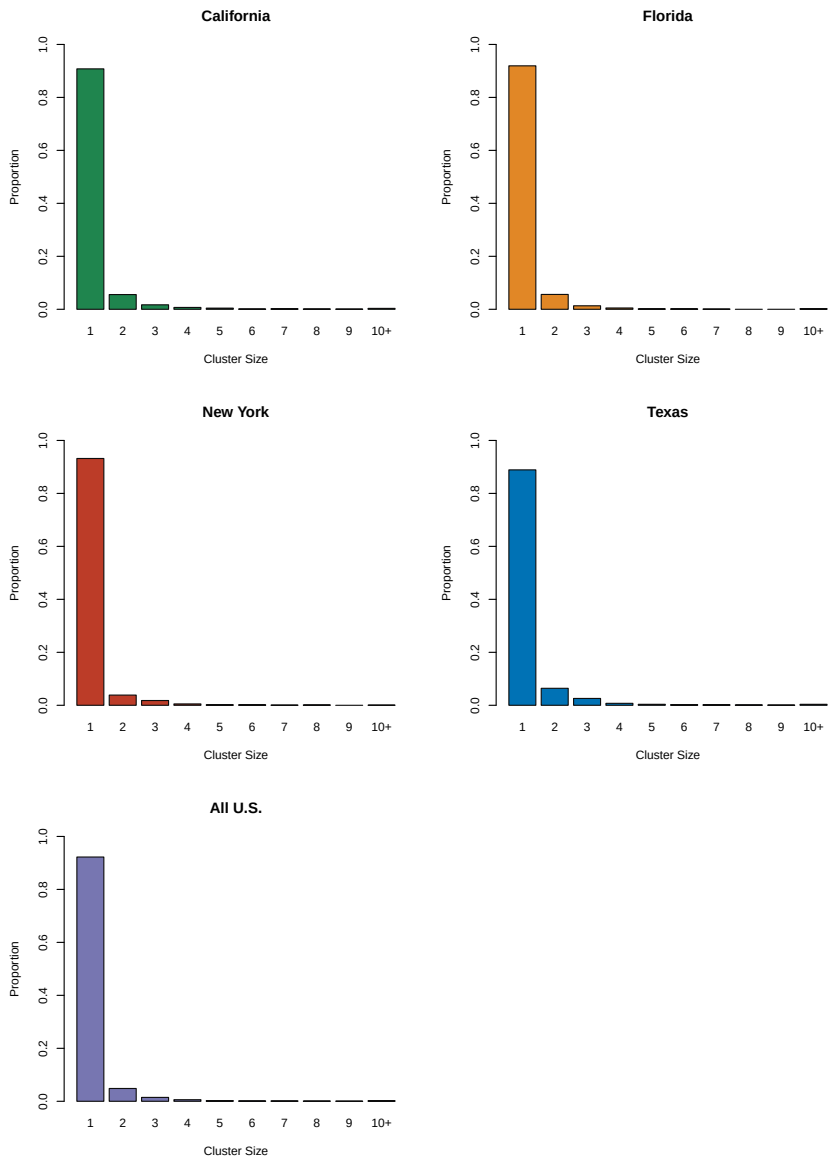

Figure 5.1: Transmission cluster size distributions in the U.S. states of California, Florida, New York, and Texas. Transmission clusters were defined at the county level as described in the methods.

State	$\hat{R}(95\%CI)$	$\hat{k}(95\%CI)$
California	0.17 (0.16-0.19)	0.09 (0.08-0.10)
Florida	0.14 (0.12-0.17)	0.09 (0.07-0.12)
New York	0.11 (0.10-0.14)	0.08 (0.06-0.11)
Texas	0.22 (0.20-0.25)	0.11 (0.09-0.13)

Table 5.2: Maximum likelihood estimates of R and k, by state

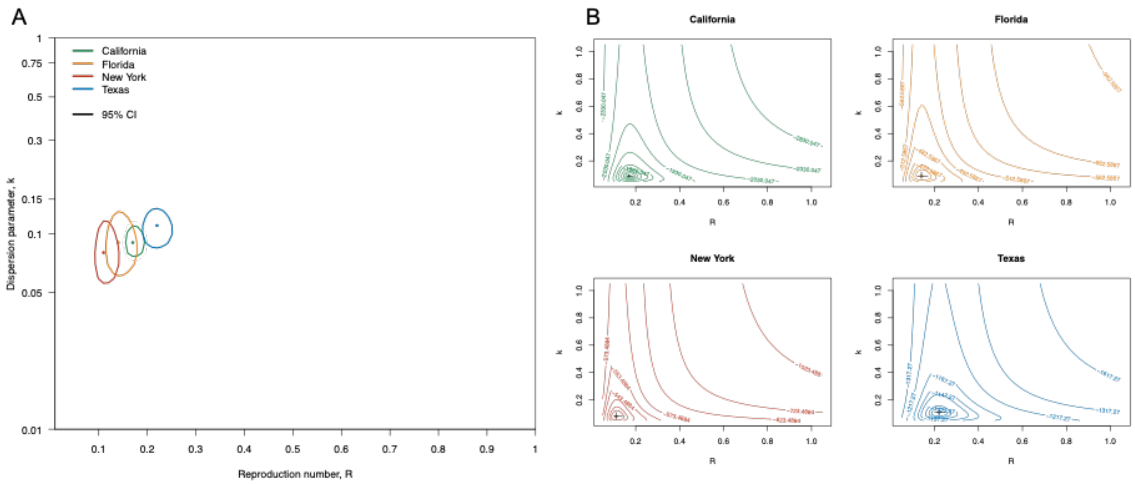

Figure 5.2: Joint estimates of R and k by state. A) Surface plots for point estimates of R and k and 95% CIs using a negative binomial branching process model. Note the y-axis is on the logarithmic scale. B) Contour plots of log-likelihood surfaces. Horizontal and vertical black lines represent 95% CIs for $\hat{R}$ and $\hat{k}$, respectively.

State	p_t, Percent (95% CI				
	80%	85%	90%	95%	100%
California	5.7 (5.3-6.1)	6.6 (6.2-6.9)	7.4 (7.0-7.8)	8.3 (7.9-8.6)	9.1 (8.7-9.5)
Florida	5.3 (4.6-6.1)	6.0 (5.3-6.8)	6.7 (6.0-7.5)	7.4 (6.7-8.2)	8.1 (7.4-8.9)
New York	4.5 (3.9-5.1)	5.0 (4.4-5.7)	5.6 (5.0-6.2)	6.1 (5.5-6.8)	6.7 (6.1-7.3)
Texas	7.0 (6.1-7.7)	8.1 (7.2-8.8)	9.2 (8.3-9.9)	10.3 (9.4-11.0)	11.4 (10.5-12.1)

Table 5.3: Expected percent of transmission attributed to a given proportion of the cases, p_t

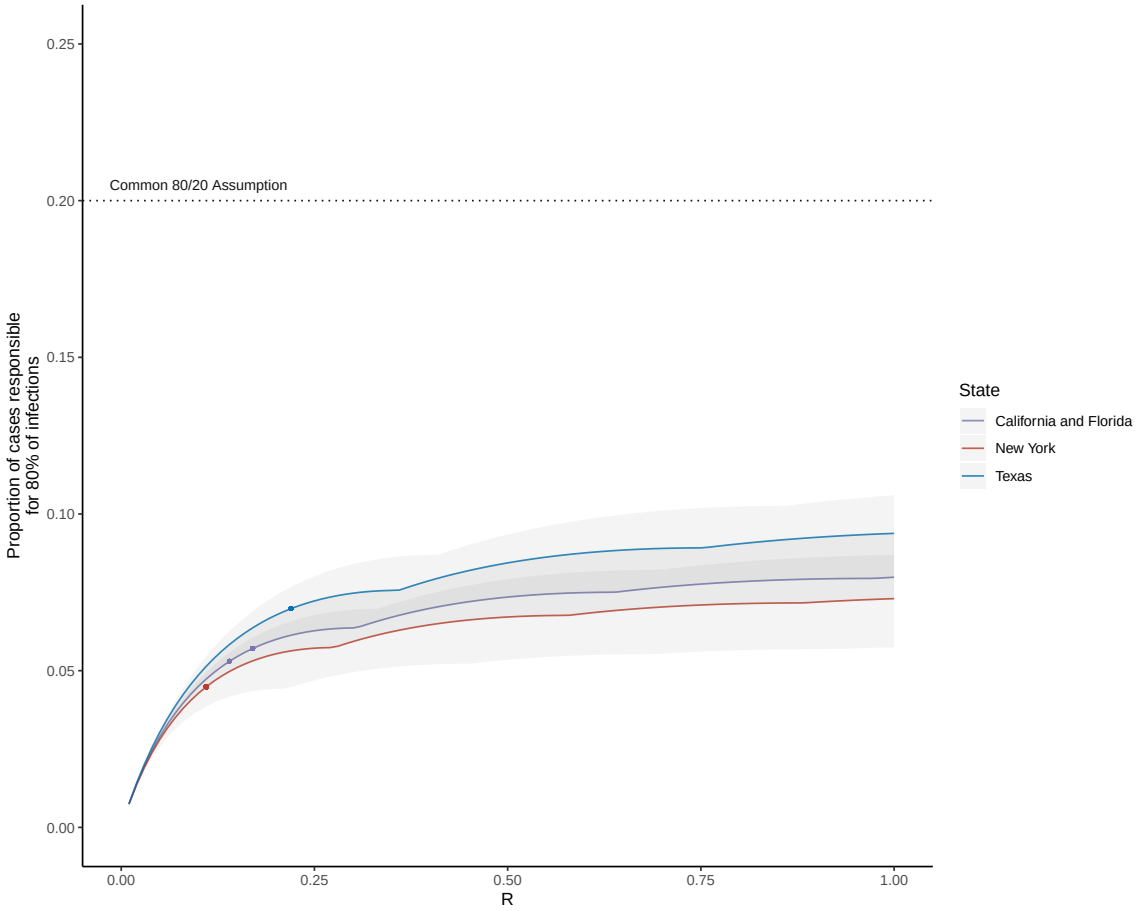

Figure 5.3: Proportion of infected individuals responsible for 80% of the total secondary transmissions (p_{80}) across the current consensus range of R values for TB in the United States. Colors lines indicate expected proportion based on MLE estimates of k (California and Florida: $\hat{k} = 0.09$; New York: $\hat{k} = 0.08$; Texas: $\hat{k} = 0.11$). Shaded areas represent 95% confidence intervals. Dots represent MLE estimates of R for the four states (California: $\hat{R} = 0.17$; Florida: $\hat{R} = 0.14$; New York: $\hat{R} = 0.11$; Texas: $\hat{R} = 0.22$).

5.6 Supplemental Materials

County Level				
Cluster Size	California	Florida	New York	Texas
1	3770	1115	1451	1927
2	231	68	60	139
3	70	16	28	56
4	29	6	8	16
5	18	2	3	8
6	5	2	3	5
7	8	1	1	4
8	5	0	2	3
9	3	0	0	2
10	3	1	0	0
11+	11	2	1	8
Total Clusters	4135	1213	1557	2168
State Level				
Cluster Size	California	Florida	New York	Texas
1	2908	883	1328	1927
2	274	95	80	139
3	88	27	25	56
4	47	10	18	16
5	19	4	2	8
6	17	3	4	5
7	15	4	3	4
8	6	2	1	3
9	14	2	0	2
10	4	0	0	0
11+	23	6	4	8
Total Clusters	3415	1036	1456	1751

Table S5.1: Cluster size distributions of TB in the U.S., by cluster definitions

State	$\hat{R}(95\%CI)$	$\hat{k}(95\%CI)$
California	0.32 (0.30-0.35)	0.13 (0.12-0.15)
Florida	0.27 (0.23-0.31)	0.17 (0.13-0.23)
New York	0.17 (0.14-0.20)	0.10 (0.08-0.14)
Texas	0.37 (0.33-0.41)	0.15 (0.13-0.19)

Table S5.2: Maximum likelihood estimates of R and k, by state, with transmission clusters defined at the state level

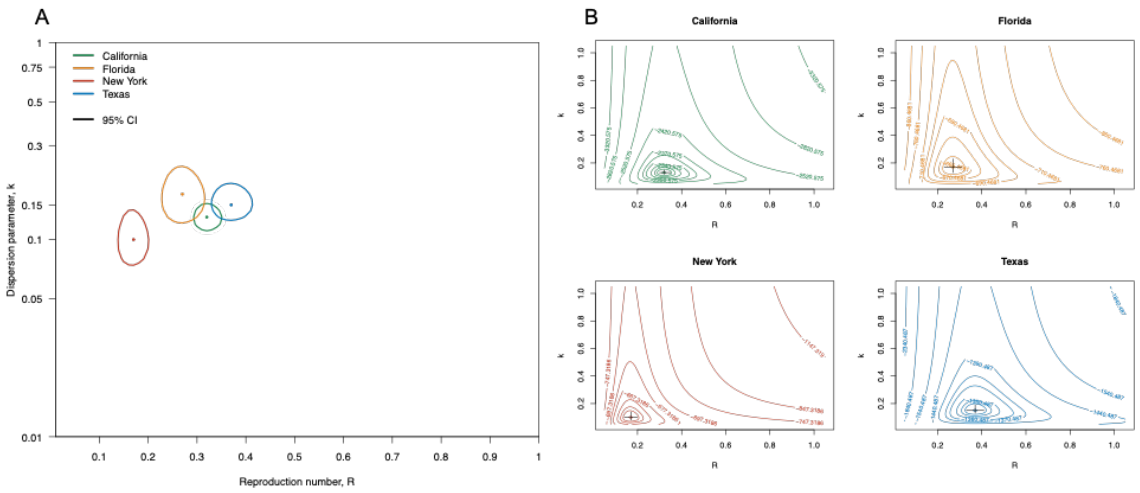

Figure S5.1: Joint estimates of R and k by state, with clusters defined at the state level. A) Surface plots for estimates of R and k using a negative binomial branching process model. B) Contour plots of log-likelihood surfaces. Horizontal and vertical black lines represent 95% of $\hat{R}$ and $\hat{k}$, respectively.

State	p_t, Percent (95% CI				
	80%	85%	90%	95%	100%
California	8.5 (8.0-9.3)	10.1 (9.6-10.9)	11.7 (11.2-12.5)	13.3 (12.8-14.1)	14.9 (14.4-15.7)
Florida	9.5 (8.2-11.0)	10.9 (9.5-12.3)	12.2 (10.9-13.7)	13.6 (12.2-16.4)	14.9 (13.6-16.4)
New York	6.1 (5.3-7.1)	6.9 (6.2-8.0)	7.8 (7.0-8.8)	8.6 (7.9-9.7)	9.5 (8.7-10.5)
Texas	9.6 (8.7-11.2)	11.5 (10.5-13.0)	13.3 (12.4-14.9)	15.2 (14.2-16.7)	17.0 (16.1-18.6)

Table S5.3: Expected percent of transmission attributed to a given proportion of the cases, p_t (Cluster definitions defined at state level)

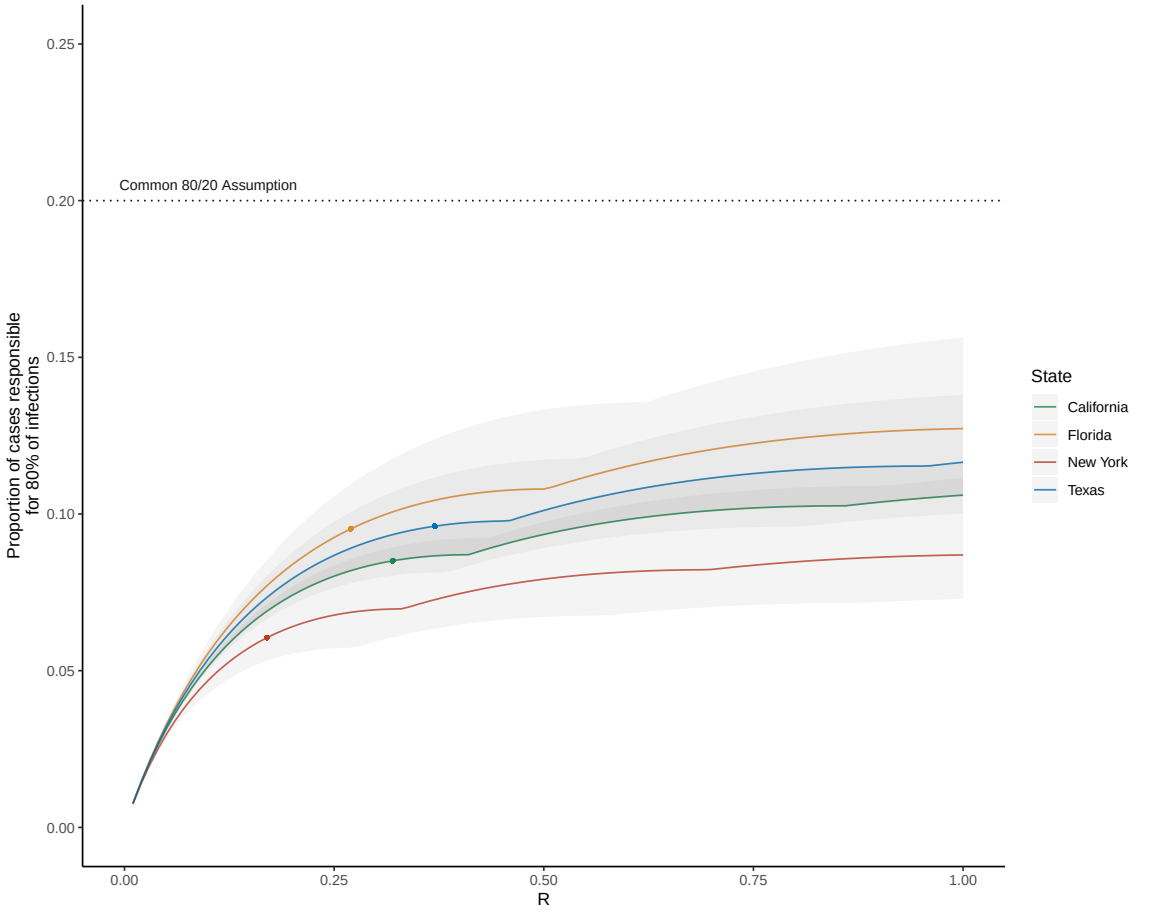

Figure S5.2: Proportion of infected individuals responsible for 80% of the total secondary transmissions (p_{80}) across the current consensus range of R values for TB in the United States, with transmission clusters defined at the state level. Colors lines indicate expected proportion based on MLE estimates of k (California: $\hat{k} = 0.13$; Florida: $\hat{k} = 0.17$; New York: $\hat{k} = 0.10$; Texas: $\hat{k} = 0.15$). Shaded areas represent 95% confidence intervals. Dots represent MLE estimates of R for the four states (California: $\hat{R} = 0.32$; Florida: $\hat{R} = 0.27$; New York: $\hat{R} = 0.17$; Texas: $\hat{R} = 0.37$).

5.7 Chapter 5 References

1. Langer AJ, Navin TR, Winston CA, et al. Epidemiology of Tuberculosis in the United States. Clinics in Chest Medicine 2019;40(4):693-702.

2. Dowdle W. A strategic plan for the elimination of tuberculosis in the United States. MMWR Morbidity and mortality weekly report 1989;38:1-25.

3. Schwartz NG, Price SF, Pratt RH, et al. Tuberculosis - United States, 2019. MMWR Morbidity and mortality weekly report 2020;69(11):286-9.

4. Haddad M, B. , Mitruka M, B. , Oeltmann J, et al. Characteristics of Tuberculosis Cases that Started Outbreaks in the United States, 2002–2011. Emerging Infectious Disease 2015;21(3):508.

5. Mitruka K, Oeltmann J, Ijaz K, et al. Tuberculosis Outbreak Investigations in the United States, 2002–2008. Emerging Infectious Disease Journal 2011;17(3):425.

6. Lloyd-Smith JO, Schreiber SJ, Kopp PE, et al. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438(7066):355-9.

7. Stein RA. Super-spreaders in infectious diseases. Int J Infect Dis 2011;15(8):e510-3.

8. Lloyd-Smith JO. Maximum Likelihood Estimation of the Negative Binomial Dispersion Parameter for Highly Overdispersed Data, with Applications to Infectious Diseases. PLOS ONE 2007;2(2):e180.

9. Zhang Y, Li Y, Wang L, et al. Evaluating Transmission Heterogeneity and Super-Spreading Event of COVID-19 in a Metropolis of China. International Journal of Environmental Research and Public Health 2020.

10. Anderson RM, Fraser C, Ghani AC, et al. Epidemiology, transmission dynamics and control of SARS: the 2002-2003 epidemic. Philosophical transactions of the Royal Society of London Series B, Biological sciences 2004;359(1447):1091-105.

11. Shen Z, Ning F, Zhou W, et al. Superspreading SARS Events, Beijing, 2003. Emerging Infectious Diseases 2004;10(2):256-60.

12. Wong G, Liu W, Liu Y, et al. MERS, SARS, and Ebola: The Role of Super-Spreaders in Infectious Disease. Cell Host and Microbe 2015;18(4):398-401.

13. Ypma RJ, Altes HK, van Soolingen D, et al. A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. Epidemiology (Cambridge, Mass) 2013;24(3):395-400.

14. Melsew YA, Gambhir M, Cheng AC, et al. The role of super-spreading events in Mycobacterium tuberculosis transmission: evidence from contact tracing. BMC Infectious Diseases 2019;19(1):244.

15. Becker N. On parametric estimation for mortal branching processes. Biometrika 1974;61(2):393-9.

16. Farrington CP, Kanaan MN, Gay NJ. Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics 2003;4(2):279-95.

17. Woolhouse MEJ, Dye C, Etard JF, et al. Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. Proceedings of the National Academy of Sciences 1997;94(1):338.

18. Severe acute respiratory syndrome–Singapore, 2003. MMWR Morbidity and mortality weekly report 2003;52(18):405-11.

19. McCreesh N, White RG. An explanation for the low proportion of tuberculosis that results from transmission between household and known social contacts. Scientific Reports 2018;8(1):5382.

20. Getz WM, Lloyd-Smith JO, Cross PC, et al. Modeling the invasion and spread of contagious diseases in heterogeneous populations. Presented at Disease Evolution: Models, Concepts, and Data Analyses2006.

21. Ferguson NM, Laydon D, Nedjati-Gilani G, et al. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Imperial College London, 2020.

Chapter 6

Summary and Conclusions

6.1 Overview

Although inter-individual heterogeneity in transmission is known to be a critical factor shaping the epidemiology of many infectious diseases, its impact on TB transmission remains largely unknown. A growing body of epidemiological work investigating individual differences in TB transmission suggest that extreme inter-individual heterogeneity and superspreading is a defining feature of TB transmission. The three studies in this dissertation have contributed additional information to help clarify the role and impact of inter-individual heterogeneity in TB transmission dynamics.

The first study developed a method to quantify the propensity for superspreading using readily available transmission cluster data in a surveillance system. Simulated data were used to test the robustness of the inference procedure and found that inference of the negative binomial parameters R and k using transmission cluster data is reliable and accurate despite several well-known limitations in imperfect surveillance. When applied to surveillance data in the United States, the results suggested a high degree of inter-individual heterogeneity. These results are in line with other studies seeking to quantify individual heterogeneity, and is plausible from a public health

standpoint as heterogeneity typically increases as public health systems improve.[1–3] This study provides the first known estimates of inter-individual heterogeneity in the U.S.

The second study was the first study to systematically quantify inter-individual heterogeneity in TB transmission using data from various global contexts. The results of this study indicate that inter-individual heterogeneity is largely consistent across populations and settings, and TB is near-universally characterized by a high propensity for superspreading. The study also demonstrated that incorporating inter-individual heterogeneity in TB transmission models could more accurately represent observed patterns of TB transmission by virtue of the relative probabilities in obtaining the largest cluster size in the observed data.

The third study was the first to quantify inter-individual heterogeneity in TB transmission in the four U.S. states responsible for the majority of TB burden. Using TB cluster data provided by the U.S. Centers for Disease Control and Prevention, results indicated a high propensity for superspreading and marked similarity of between states. These results also held when expanding the definition of transmission cluster to provide a more conservative definition. The study further demonstrated a small minority of cases ($\sim$7-11 percent) were responsible for all secondary TB transmission in the population.

Taken together, the findings from this dissertation align with more recent studies suggesting that TB transmission is characterized by marked variability in individual transmission.[1,4,5] Moreover, the evidence base formed from this body of work suggests that individual variation in TB transmission may be equal to or more extreme than other infectious diseases well-known to be characterized by superspreading, including Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV; k=0.16),[6] Middle East Respiratory Syndrome Coronavirus (MERS-CoV; k=0.26),[7] COVID-19 (SARS-CoV-2, k=0.10),[8] monkeypox (k=0.33),[9] and Ebola virus (k=0.09).[10] As a result, public

health programs seeking to mitigate TB transmission may draw from successes of these and other diseases with similar heterogeneity.

6.2 Public Health and Epidemiological Implications

The accurate representation of inter-individual heterogeneity in TB transmission continues to be a challenge in TB modeling. This dissertation's goal is primarily to benefit future epidemiologic modeling efforts by accounting for the diversity of unknown and unmeasured factors (and the interaction between them) that contribute to variation in the number of secondary cases. It accomplishes this goal by both inferring the first empirical estimates of this parameter globally, and by providing modelers with a tool to estimate the propensity for superspreading within their study population. This dissertation also highlights the need to report estimates of inter-individual heterogeneity and provides a common reference (by virtue of k) for which to compare populations and diseases. By incorporating such previously unknown information into epidemic models, a more accurate representation of epidemic spread and the impact of intervention efforts can be realized. Although this dissertation focused intensely on one aspect in the broader context of epidemiologic heterogeneity, the importance of quantifying inter-individual heterogeneity cannot be overstated and its findings will prove critical in better informing intervention efforts, decision making, and the allocation of resources.

This dissertation provided preliminary evidence that the vast majority of incident TB does not lead to secondary outbreaks, yet those that do often result in explosive outbreaks that account for the majority of secondary transmission. This is the defining hallmark of other diseases known to be characterized by extensive superspreading.[4] Such information has direct implications for improving the effectiveness and efficiency of outbreak response programs. First, it demonstrates that the presence of at least

one secondary case arising due to recent transmission greatly increases the likelihood of a large outbreak. Thus, outbreak response teams may need to respond earlier and with more urgency to prevent additional ongoing TB cases that may have occurred under the current response strategy.

6.3 Limitations

The methods proposed in this dissertation rely on several assumptions that should be considered. First, the type of branching process used in this dissertation, a Galton-Watson process, assumes that the depletion of susceptible contacts due to infection is negligible and thus there is an infinite susceptible population. While this assumption likely holds true in the general population of large, low-incidence settings such as the United States, it may not be as reliable in smaller populations with a high prevalence of TB or in vulnerable communities in low-incidence settings. Additionally, transmission and subsequent disease spread is assumed to be independent and identically distributed (i.i.d.). This assumption may be violated if there are correlations between source cases and contacts, such as behavior or clustering of highly susceptible individuals. Additionally, the number of observed secondary cases is drawn from a negative binomial distribution. While this distribution allows for an unknown degree of heterogeneity, it may not be definitively superior to other distributions. For instance, a recent study employed the use of a Poisson log-normal (PLN) distribution in comparison to the negative binomial.[11] Whereas the negative binomial is a mixture of a gamma-distributed λ in a Poisson process, the offspring distribution follows a PLN distribution if $log(\lambda)$ is normally distributed. In their study, the PLN models better captured the long tails of the distribution of cluster-level data and were statistically better fits than the negative binomial. However, cluster data analyzed genetic clusters and the long tails are likely a product of multiple transmission clusters. Moreover,

the model did not account for overlapping clusters or censorship.

While the simulations in the methods developed in Study 1 demonstrate that the inference procedure is robust under perfect surveillance, the transition from theoretical evaluation to real world application also poses several limitations. All three studies applied these methods to real-world surveillance data containing imperfect transmission clusters, missing cases, and censoring. Although these limitations were evaluated in the method developed in Study 1, those evaluations specified the degree to which these imperfect surveillance measures impacted estimates of heterogeneity. In reality, these factors are unknown to surveillance systems and thus the degree of bias introduced is similarly unknown. Such issues present a larger problem in limited-resource settings where surveillance systems inadequately capture information on new TB cases.

Perhaps most importantly, this approach is of primary use when evaluating recent transmission and does not estimate complete heterogeneity in TB transmission, which may include infected individuals that result in latent TB infection (LTBI). Unique to the natural history of TB, it is possible that certain index cases may be prolific transmitters of successful TB infection yet result in little or no secondary cases. Unfortunately, there is little epidemiologic work investigating the relationship between secondary infections and secondary cases caused by an individual. To our knowledge, only one well-designed study directly compared the number of infections with the number of resultant cases and found a high degree of concordance.[6] Though this indicates that those who infect more also result in more secondary cases, there is too little evidence for scientific consensus.

6.4 Remaining Gaps in Knowledge and Future Directions

While this body of work provides a preliminary framework for assessing inter-individual heterogeneity in TB transmission, additional research is greatly needed to obtain a more complete understanding of TB transmission dynamics on an individual level. Many TB researchers reasonably argue that the plateauing rate of decline in global TB incidence is more a function of reactivation of LTBI than recent transmission.[12–14] Thus, a more complete understanding of inter-individual heterogeneity in TB transmission – including the number of both secondary cases and secondary infections leading to LTBI – would prove useful in identifying recently infected contacts who would benefit from preventive therapy and allocating resources. This may be achieved in the context of a multi-type branching processes model that accounts for LTBI or through other novel means of epidemiologic modeling. However, given challenges associated with identifying the source of infection for individuals with LTBI, there is little data from which models may draw assumptions in this regard. Further epidemiological research is needed to shed light on this unique and critical aspect of TB transmission.

Our definition of inter-individual heterogeneity accounted for all known and unknown factors in the infectious history of both the source and the contact. While this provided a benefit to the methods presented here aimed at quantifying this heterogeneity, public health systems benefit from being able to proactively identify tangible situations with a higher potential for superspreading. Despite a large body of work contextualizing risk factors for individual infectiousness and susceptibility, these are often discussed separately and remain limited in their ability to capture the full extent of factors that lead to superspreading.[15] However, analysis of risk factors associated with transmission clusters as a whole would jointly describe the complex dynamic

of host and contact factors that resulted in extensive transmission, and afford the ability to predict superspreader-associated clusters. Drawing on previous models and analysis of TB transmission dynamics, the methods and evaluation presented in this dissertation afford the opportunity for *de novo* models of TB outbreaks that more accurately reflect observed TB transmission patterns. As a direct result of this dissertation, novel TB outbreak models can be developed using an individual-based framework that more accurately accounts for unknown factors attributing to superspreading.

1. Ypma RJ, Altes HK, van Soolingen D, Wallinga J, van Ballegooijen WM. A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. Epidemiology (Cambridge, Mass) 2013;24:395-400.

2. McCreesh N, White RG. An explanation for the low proportion of tuberculosis that results from transmission between household and known social contacts. Scientific Reports 2018;8:5382.

3. Trauer JM, Dodd PJ, Gomes MGM, et al. The Importance of Heterogeneity to the Epidemiology of Tuberculosis. Clinical Infectious Diseases 2018;69:159-66.

4. Melsew YA, Gambhir M, Cheng AC, et al. The role of super-spreading events in Mycobacterium tuberculosis transmission: evidence from contact tracing. BMC Infectious Diseases 2019;19:244.

5. Gardy JL, Johnston JC, Sui SJH, et al. Whole-Genome Sequencing and Social-Network Analysis of a Tuberculosis Outbreak. New England Journal of Medicine 2011;364:730-9.

6. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438:355-9.

7. Kucharski AJ, Althaus CL. The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. Eurosurveillance 2015;20:21167.

8. Endo A, null n, Abbott S, Kucharski A, Funk S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China [version 3; peer review: 2 approved]. Wellcome Open Research 2020;5.

9. Blumberg S, Lloyd-Smith JO. Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. PLoS Comput Biol 2013;9:e1002993.

10. Toth DJA, Gundlapalli AV, Khader K, et al. Estimates of Outbreak Risk from New Introductions of Ebola with Immediate and Delayed Transmission Control. Emerging infectious diseases 2015;21:1402-8.

11. Brooks-Pollock E, Danon L, Korthals Altes H, et al. A model of tuberculosis clus-

tering in low incidence countries reveals more transmission in the United Kingdom than the Netherlands between 2010 and 2015. PLoS Comput Biol 2020;16:e1007687-e.

12. Mancuso JD, Diffenderfer JM, Ghassemieh BJ, Horne DJ, Kao TC. The Prevalence of Latent Tuberculosis Infection in the United States. American journal of respiratory and critical care medicine 2016;194:501-9.

13. Shrestha S, Cherng S, Hill AN, et al. Impact and Effectiveness of State-Level Tuberculosis Interventions in California, Florida, New York, and Texas: A Model-Based Analysis. American journal of epidemiology 2019;188:1733-41.

14. Yuen CM, Kammerer JS, Marks K, Navin TR, France AM. Recent Transmission of Tuberculosis — United States, 2011–2014. PLOS ONE 2016;11:e0153728.

15. Haddad M, B. , Mitruka M, B. , Oeltmann J, Johnson WD, Jr., Navin TR. Characteristics of Tuberculosis Cases that Started Outbreaks in the United States, 2002–2011. Emerging Infectious Disease 2015;21:508.

Appendix A

Key Formulas

A.1 Probability Density Function

The probability of a transmission chain originating with n index cases resulting in a final transmission cluster size of y can be expressed as:

$$P(Y = y|n) = \binom{n}{y} \frac{\Gamma(ky + y - n)}{\Gamma(ky)\Gamma(y - n - 1)} \frac{(\frac{R}{k})^{y-n}}{(1 + \frac{R}{k})^{ky+y-n}}$$

A.2 Likelihood Equation

The likelihood of parameters R and k with data containing $A_{y,n}$ fully observed clusters resulting in size y and initiating with n index cases and $B_{y,n}$ censored clusters of size y with n index cases is"

$$L\left(R, k|\vec{A}, \vec{B}\right) = \prod_{y_a}\prod_{n_a} P(Y = y|n)^{a_{y,n}} \prod_{y_b}\prod_{n_b} P(Y \geq y|n)^{b_{y,n}}$$

Where $P(Y \geq y|n) = 1 - \sum_{i=1}^{y-1} P(Y = y|n)$

Appendix B

Relevant R Code for Inference Procedure

B.1 Branching Process Function

```
bp <- function(gens=20, init.size=1, offspring, ...){
  Z <- list() #initiate the list
  Z[[1]] <- init.size
  i <- 1
  while(sum(Z[[i]]) > 0 && i <= gens) {
    Z[[i+1]] <- offspring(sum(Z[[i]]), ...)
    i <- i+1
  }
  return(Z)
}
```

B.2 Imperfect Simulation Function

```
##'
   ------------------------------------------------------------------

##' Simulating imperfect observation
##'      @param true_R       True underlying R value for NB
   branching process
##'      @param true_k       True underlying k value for NB
   branching process
##'      @param num_chains   Number of simulated transmission
   chains in a surveillance system
##'      @param p1           Probability of ascertaining cases
    by passive surveillance
##'      @param p2           Probability of ascertaining cases
    by active surveillance
##'      @param prob_cens    Probability that a chain will be
   censored
##'      @param perc_overlap Proportion of clusters that
   overlap (i.e. multiple index cases)
##'      - - - - - - - - - - - - - -
##'      @return Output is a list of length 2, each list
   contains a data frame of cluster sizes, index
##'              cases, and censored status, for:
##'                 [[1]] Perfect Surveillance
##'                 [[2]] Imperfect Surveillance
##'
   ------------------------------------------------------------------
```

```
##'
imperfect <- function(true_R, true_k, num_chains=2000, p1=0.75
   , p2=0.5, prob_cens=0.10, perc_overlap=0.20){
    z <- replicate(num_chains,bp(offspring = rnbinom, mu =
       true_R, size = true_k)) # Simulate individual-level
       surveillance system
    z.pass <- z; z.act <- z
   ## - - - - - - - - - - - - - - - - - - - - - -
   ## Imperfect case ascertainment
   ## - - - - - - - - - - - - - - - - - - - - - -
   #Passive surveillance
    for (i in 1:length(z.pass)){
      for (j in 1:length(z.pass[[i]])){
        for (k in 1:length(z.pass[[i]][[j]])){
          for (l in 1:length(z.pass[[i]][[j]][[k]])){
            if (runif(1)<(1-p1)){                    # Get to
                "individuals" and randomly assign NA based on
               p1
              z.pass[[i]][[j]][[k]]<-NA
            }}}}}
    #Active case finding (only chains with at least one case
       observed by passive surveillance)
        ##' Note to self: Index cases with no secondary cases
           present in the branching process as a
        ##' list of 2 (always [1] and [0]). The second value
           is what we are concerned about having an NA value
           via passive
        ##' surveillance. In the scenario where the list is [1
           ] and [NA], this would artifically make the chain
```

```
        eligible for
    ##' reevaluation since the cluster is not completely
        missing. So before reevaluation with p2 we need to
    ##' assign the first position in all of the lists as
        NA. That way, if the index case is missing,
    ##' both values will be missing.
a<-z.pass
for (i in 1:length(a)){
  a[[i]][[1]]<-NA
}
b<-cbind(a,sapply(a, function(x) all(is.na(unlist(x))))) #
    Determine if anyone in the chain has been seen
for (i in 1:length(z.pass)){
  if (b[i,2]==TRUE){
    z.act[[i]]<-a[[i]] # Skip if cluster not observed at
        all
  } else {
    for (j in 1:length(z.pass[[i]])){
      for (k in 1:length(z.pass[[i]][[j]])){
        for (l in 1:length(z.pass[[i]][[j]][[k]])){
          if (is.na(z.pass[[i]][[j]][[k]])){
            if(runif(1)<=(p2)){       #Active probability
               of being seen by case detection
              z.act[[i]][[j]][[k]]<-z[[i]][[j]][[k]] #
                 Reassign original value
            } else {z.act[[i]][[j]][[k]]<-z.pass[[i]][[j
               ]][[k]]}
          }}}}}
# "Break" chains based on the position of missing cases in
```

```
    the chain
l <- z.act  #dummy/temp data to not change z.act
for (i in 1:length(l)){
  l[[i]][[1]]<-NULL #remove first position of the nested
     list so that it eases summing lengths (can't sum
     based on integer values in imperfect observations)
}
#"Break apart" the chains
t1 <- lapply(lapply(seq_along(l), function(nm) {split(l[[
   nm]], cumsum(sapply(l[[nm]], function(x) all(is.na(x)))
   ))}), function(lstA) lapply(lstA,function(x) Filter(
   function(y) !all(is.na(y)), x)))
t2 <- rapply(unlist(t1,recursive=FALSE),function(x) x[!is.
   na(x)], how="replace") #Remove NA values.
z.broken <- Filter(length,t2) #remove all with length 0 (
   missing/unobserved)
## - - - - - - - - - - - - - - - - - - - - -
## Censoring
## - - - - - - - - - - - - - - - - - - - - -
z.cen <- z.broken                          # Initialize
   the censored list
for (i in 1:length(z.broken)){             # Iterate
   through the list
  if (length(z.broken[[i]])>1) {           # List must
     have at least length of two (cant be censored if the
     index case isnt seen, then it is unobserved as above)
    if(runif(1)<=prob_cens){               # Stochastic
       process to determine if the nested list will be
       censored
```

```
      if(length(z.broken[[i]])==2){
        n <- 2} else {                    # A little
           trick to get over the issue of sample(2:2,1)
           returning values of 1 as well as 2
        n <- sample(2:length(z.broken[[i]]),1)}   #
           Randomly determine what list position in the
           nested list will be the censor threshold
      z.cen[[i]][n:length(z.broken[[i]])] <- NA   # Fill
         all positions from n to the end of the nested
         list with NA
    }}}
out_list <- lapply(z.cen, function(x) {    # Remove all
   nested list elements that contain NA
                  inds <- sapply(x, function(x) any(is.na
                     (x)))
                  if(any(inds)) x[seq_len(which.max(inds)
                     - 1)] else x})
cens <- numeric(length(out_list))
true <- numeric(length(out_list))
for (k in 1:length(out_list)){
  cens[k]<-sum(lengths(out_list[[k]])) # Get cluster size
     of censored clusters
  true[k]<-sum(lengths(z.broken[[k]])) # Get cluster size
     of uncensored (but imperfect obs) clusters
}
Y_cens <- data.frame(y.cens=cens, censor=ifelse(cens!=true
   ,1,0)) #Create a censoring index (1=censored, 0=
   uncensored)
## - - - - - - - - - - - - - - - - - - - - - - -
```

```
## Overlapping clusters
## - - - - - - - - - - - - - - - - - - - - - -
#' Determine sampling space - i.e. how many clusters get
   merged with each iteration
#'     Skewed towards smaller overlapping clusters (n=2,3)
   , but allows for up
#'     to 7 (~0.02% chance) based on Poisson w/ lambda 1
a <- data.frame(table(rpois(1000000,1))) # Drawn from a
   Poisson with lamda=1
a_trunc <- a[a$Var1 %in% c(2:7),]        # Restrict the
   number of overlapping clusters (n) to between 2-7,
   heavily skewed towards lower values
n <- round(nrow(Y_cens)*perc_overlap)          # Determine
   the number to be merged
sample_clust <- sample(c(rep(2,a_trunc[1,2]), # ~184000 or
    ~69%
                         rep(3,a_trunc[2,2]), # ~61200  or
                             ~23%
                         rep(4,a_trunc[3,2]), # ~15100  or
                             ~6%
                         rep(5,a_trunc[4,2]), # ~3000   or
                             ~1%
                         rep(6,a_trunc[5,2]), # ~500    or
                             ~0.2%
                         rep(7,a_trunc[6,2])),# ~60     or
                             ~0.02%
                       size=n, replace=TRUE)
names(sample_clust) <- paste0("S", 1:n)
m <- nrow(Y_cens)-sum(sample_clust) # Determine the number
```

```
    that will not be merged (m)
non_merge_clust <- rep(1, m)        # Create a vector with
    replicated 1 based on m
names(non_merge_clust) <- paste0("N", 1:m)
# Combine sample_clust and non_merge_clust, and then
   randomly sort the vector
combine_clust <- c(sample_clust, non_merge_clust)
combine_clust2 <- sample(combine_clust, size=length(
   combine_clust))
# Expand the vector
expand_list <- list(lengths=combine_clust2, values=names(
   combine_clust2))
expand_clust <- inverse.rle(expand_list)
# Create a data frame with y and expand_clust
dat <- data.frame(Y_cens, group=factor(expand_clust,
   levels=unique(expand_clust)))
dat$index <- 1 # add the index case number for summing
# Convert dat2 to a matrix, sum the index cases and
   censoring index, remove the group column
dat2 <- aggregate(cbind(dat$y.cens, dat$index, dat$censor)
   , by=list(group=dat$group),FUN=sum)
dat2$group <- NULL
y.merged <- as.matrix(dat2); colnames(y.merged)<-c("clust_
   size","index_cases","censor_status")
y.final <- data.frame(y.merged) #just to be safe
  y.final$censor_status <- ifelse(y.final$censor_status>=1
     ,1,0) # if more than 1 censored clusters merged
## - - - - - - - - - - - - - - - - - - - - - - - - - - - -
    - - - - - - - - - - - - - - -
```

```
    y.true <- unlist(lapply(z,function(x) sum(unlist(x)))) #
       Sum true cluster sizes
    Y.true <- data.frame(y.true=y.true,  index=rep(1,times=
       length(y.true)), censor=rep(0,times=length(y.true)))  #
        original uncensored and unmerged data
      names(Y.true) <- c("clust_size","index_cases","censor_
         status")
    return(list(Y.true, y.final))
      #return(list(z, z.pass, z.act, z.broken, out_list, cens,
          true, Y_cens, y.final)) #for validation
}
```

B.3 Likelihood Function

```
cens_likelihood <- function(Y,R,k) {
  p_function <- function(y,n){          #Dummy function to
     apply
    exp(log(n)-log(y)+lgamma(k*y+y-n)-(lgamma(k*y)+lgamma(y-n+
       1))+(y-n)*log(R/k)-(k*y+y-n)*log(1+R/k)) #PDF as
       defined un methods
  }
  ya <- Y[Y[,3]==0,] # Uncensored clusters
  yb <- Y[Y[,3]==1,] # Censored clusters

  liks_a <- log(p_function(ya[,1],ya[,2])) # Can apply P(Y=y)
     via vectorization

  liks_b <- numeric(nrow(yb))             # Not sure how to
     vectorize with the $P(Y \geq y)$ being of the 1-sum(p_
     function(1:(y-1),n)) below
  if(nrow(yb)>0){                         # This for loop
     approach is reasonably fast (about 9 seconds on a list of
      2000 cluster sizes)
    for (i in 1:nrow(yb)){
      y <- yb[i,1]
      n <- yb[i,2]
      if (y==1){                         # If the cluster size
         is 1, the P(1)=1, thus log(1)=0
        liks_b[i] <- 0                     # This trick
           prevents issues with running the code
      } else{
```

```
        if (is.nan(log(max(10^-300,1-sum(p_function(1:(y-1),n)
           ))))){ # Trick to avoid NaN due to extremely
           unlikely clusters (very rare, but was causing
           numeric overflow problems)
          liks_b[i] <- log(10^-300)
        } else {liks_b[i] <- log(max(10^-300,1-sum(p_function(
           1:(y-1),n))))}
      }}}
  sumliks <- sum(liks_a,liks_b)
  return(sumliks)
    #return(list(liks_a,liks_b)) #validate
}
```

B.4 Parameter Estimation Function

```
##'
   ------------------------------------------------------------------------

##' Parameter Estimation
##' Estimates MLE and confidence interval for R and k
##'      @param simdata 3-column data frame or matrix
   containing
##'                          [1] Custer Size
##'                          [2] Index Cases
##'                          [3] Censored status
##'      @param Rrange Range of R values for optimization
##'      @param krange Range of k values for optimization
##'      @param conf.interval Desired confidence interval (as
   decimal, i.e. 0.95)
##'      @param k_only Option to only estimate k values (
   increases speed significantly)
##'      - - - - - - - - - - - - - -
##'      @return Matrix containing point, lower, and upper
   bound estimates for R and k
##'
   ------------------------------------------------------------------------

paramests <- function(simdata, Rrange, krange, conf.interval=0
   .95, k_only=FALSE){
  if (k_only==TRUE){
    Rrange_2 <- 1-(mean(simdata[,2])/mean(simdata[,1])) #R MLE
        value
```

```
} else {Rrange_2 <- Rrange}
likesurf <- matrix(NA, nrow=length(Rrange_2),length(krange))
for(i in 1:length(Rrange_2)){
  for(j in 1:length(krange)){
    likesurf[i,j] <- cens_likelihood(simdata,Rrange_2[i],
       krange[j])
  }
}
chiV<-qchisq(conf.interval, df=1)/2
profprep_k <- apply(likesurf,2,function(x){max(x)})
profprep_k2 <- krange[profprep_k-max(profprep_k)>-chiV]
profprep_R <- apply(likesurf,1,function(x){max(x)})
profprep_R2 <- Rrange_2[profprep_R-max(profprep_R)>-chiV]
likesurf_max <- likesurf==max(likesurf)

output <- matrix(NA,2,3)
output[1,1] <- Rrange_2[sum(seq(1,length(Rrange_2))%*%
   likesurf_max)] #k point estimate
output[1,2] <- min(profprep_R2) #k lower CI
output[1,3] <- max(profprep_R2) #k upper CI
output[2,1] <- krange[sum(likesurf_max%*%seq(1,length(krange
   ))))] #R point estimate
output[2,2] <- min(profprep_k2) #R lower CI
output[2,3] <- ifelse(max(profprep_k2)==max(krange),Inf,max(
   profprep_k2)) #R upper CI
colnames(output) <- c("point_est","lower_ci","upper_ci");
   rownames(output) <- c("R_hat","k_hat")
if (k_only==TRUE) {
  output <- output[2,]
```

```
    }
    return(output)
}
```