PROBABILISTIC MODELING OF GENOME EVOLUTION AND DISEASE SPREAD OF TUBERCULOSIS

By

Lei Yao

A Dissertation Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: MATHEMATICS

Examining Committee:

_____

Peter Kramer, Dissertation Adviser

_____

Kristin Bennett, Dissertation Adviser

_____

Gregor Kovacic, Member

_____

Qiang Ji, Member

Rensselaer Polytechnic Institute
Troy, New York

November 2014
(For Graduation December 2014)

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

First and foremost, I would like to thank my thesis advisers: Professor Peter Kramer and Professor Kristin Bennett. I would like to thank Prof. Bennett for giving me this opportunity to work on TB disease project and funding my study. This work is supported by NIH NLH grant R01LM009731.

I thank Peter for teaching me the fundamentals of probability theory. To me, this knowledge is not only a skill that I make living on, but also it is a tool that I understand the world. Peter also helped me with my thesis tremendously. He had shown me the subtleties when applying probability theory to model a real world problem, which I am sure I cannot learn from any textbook. Kristin provided me guidance on my work on data analysis and helped me find the linkage between mathematical model and real data. She had shown me the importance of details, which is useful in many other areas in life. I would also like to express my gratitude to other two members in my committee: Professor Qiang Ji and Professor Gregor Kovaci, for their advises and suggestions on my thesis. I am impressed by the professionalism of all four of my committee members. They all helped me to be a better researcher in mathematical science. I also thank Dr. Philip Supply from Institut Pasteur de Lille, Lille, France and Nalin Rastogi from Institut Pasteur de la Guadeloupe, Guadeloupe, France. For their valuable advice on thesis.

As a chemical engineering major in college, the transition to mathematics was not an easy path for me. I would like to thank Professor Lorenzo M. Polvani in Columbia University. It was through his course on complex variables that led me to the wonderful world of mathematics and later made the decision to pursue a PhD in this field.

Next, I would like to thank all friends at RPI: Ke Wu, Yang Li, Taoran Li, Peter Muller, Jamie Blondin. The conversations with them gives me many inspirations on research. The basketball games with them helps me keep healthy and refreshed from work. For those who are still in the process of their PhD studies, I send my best wishes.

Last but not the least, I would like to thank my families for their selfless support, both mentally and financially. My parents and my sister have contributed enormously to me. Without them, all these are impossible. I want to thank my girlfriend, Jill Gao. Her love and care helped me go through a lot of down swings in the past few years. Moreover, her excellent cooking definitely helped me stay full and healthy when I am busy writing this thesis.

ABSTRACT

Tuberculosis (TB) remains one of the leading causes of mortality worldwide. It is caused by *Mycobacterium tuberculosis* complex (MTBC). The development of the DNA fingerprinting technologies in the past decade has enriched the information available for scientific research and TB control. The genetic dissimilarities among different strains of MTBC will result in different fingerprinting data. With this information, TB patient isolates can be grouped into small clusters, which greatly facilitates TB control and surveillance. *Spacer Oligonucleotide Types* (Spoligotypes) and *Mycobacterial Interspersed Repetitive Units - Variable Number Tandem Repeats* (MIRU-VNTR) are two of the popular biomarkers used for DNA fingerprinting worldwide. MIRU-VNTR typing records the numbers of the tandem repetitive units at several specific loci in MTBC genome, which are collectively referred as MIRU. This thesis studies TB from two aspects: 1)micro-level exploring the evolution properties of MIRU based on our assumptions ; 2)macro-level, proposing mathematical model to capture the transmission dynamics within a TB cluster.

In the past decade, MIRU is gaining popularity in TB research and control. Compared to Spoligotypes, MIRU is relatively less studied. Understanding the characteristics of MIRU is crucial to fully harness its power as a TB analysis tool. In the first part of this thesis, we take advantage of the characteristics of Spoligotypes to infer the mutation directions of MTBC and analyze the mutations in MIRU. A *Markov Chain* of the repeat numbers of MIRU at each locus is built based the mutations found in the data. We compute and compare the stationary distributions of repeat numbers at different loci. An error analysis is done to investigate the errors produced at each stage of our study: from inferring the probability transition matrices of the Markov Chains to computing the corresponding stationary distributions. We also study the distance between the current distribution of the repeat numbers and the stationary ones. Finally, we analyze the rates of each locus reaching its stationary distribution through theoretical computations and simulations.

In the second part, we study the transmission dynamics of TB disease. Based

on the DNA fingerprints of the MTBC, TB patients can be clustered in to small groups. This allows us to investigate the dynamics at the individual level. Since immigrants make the majority of TB cases in the United States, we focus our analysis on immigrant TB patients. We propose a model to estimate the probability of an immigrant entering the country latently infected with TB versus he/she being infected after entry, given the entry and diagnosis time of the immigrant patients within the cluster. The transmission routes among the patients increase exponentially with the size of the cluster. The fact that individuals outside the cluster could also infect someone within the cluster further complicates the dynamics. We use *Mean Field* approximations to simplify the complicated transmission routs among patients and the effects of the individuals outside the cluster. The performance of the model is evaluated with *Receiver Operating Characteristic* (ROC) analysis on simulated data. Finally, we apply our model to the patient data collected from New York City.

CHAPTER 1
Introduction

Tuberculosis (TB) is one of the most common infectious diseases worldwide. Each year, it causes approximately 2 million deaths. It is the No. 2 cause of death due to a single infectious disease (first is HIV) [1].

1.1 The Disease

TB has been a major health problem for globally for many years [1]. The death rate has once reached 800 - 1000 per 100,000 population per year in various cities in Europe in the 19th century [2]. In the early 20th century, the incidence and fatality rate began to decline as the living standards, including medical care, personal hygiene, nutrition, housing, and etc., improved. However, TB started to reemerge in the early 1990s as the number of individuals with HIV/AIDS grew, and started to be co-infected with TB [3]. The overall burden of TB continues to rise each year as the world population rapidly grows. The WHO declared TB a global health emergency in 1993 [1]. In 2012, 8.6 million people were infected with TB and 1.3 million died from it [1].

About one-third of the world population has been infected with TB, but not yet become ill [4]. This is called the latent state. People with latent TB will not transmit the disease. A small portion of people with latent TB infection will develop active TB disease. People who have inferior immune systems, such as those with HIV have much higher risk of developing active TB disease. When his/her latent TB infection becomes active, the person will develop symptoms including cough, fever, night sweats, weight loss, etc. More importantly, patients with active TB disease can transmit the bacteria to other individuals [4].

1.2 Transmission

Tuberculosis is a contagious disease caused by *Mycobacterium tuberculosis* complex (MTBC), a small, aerobic bacterium. It mainly attacks lungs, causing

pulmonary tuberculosis. However, the TB bacteria can attack many other parts of the body such as the kidney, spine and brain, causing extrapulmonary tuberculosis [1]. Although there are two forms of TB: pulmonary and extrapulmonary, only the former is transmissible. When a person with active pulmonary TB coughs, sneezes or speaks, the bacteria is spread into the air in the droplets. People nearby may breathe in these bacteria and become infected [5].

Only a small portion of individuals that are infected develop progressive disease immediately. Most people, after their initial exposure to the MTBC, will mount an immune response which prevents the bacteria from proliferating. These individuals, although carrying TB bacteria, will not show any symptoms nor will they become infectious. This status is called latent TB infection [6]. However, it is possible that the internal immune system fails to prevent the MTBC proliferation at a later point of life. In this case, the TB bacteria will become active and the individual will become sick with TB disease. Acquiring TB in this manner is called endogenous reactivation. There is a 5 - 10% risk for endogenous reactivation to happen with latent TB infection [5]. Individuals with latent infection could also be infected by other active TB carriers again. We refer this type of transmission as exogenous infection. The patient with latent TB infection could therefore develop active TB disease through two ways: endogenous reactivation or exogenous infection [7]. The dynamics of the transmission of TB is shown in Figure 1.1.

TB occurs in every part of the world. Currently, the largest number of new cases arise in Asia, which accounts for 60% of new cases globally [1]. In 2009, 22 low and middle-income countries account for more than 80% of the active cases in the world, with the five highest prevalence countries being India and China, South Africa, Nigeria, and Indonesia [8]. In the past 5 decades, globalization contributes significantly to the spread of TB. Immigration from countries with high TB incidence to countries with low TB incidence is causing a major problem in TB spread [9]. In the most developed countries, at least 50% of the TB cases are among foreign born people [1]. In the United States, the TB incidence rate among foreign-born persons in 2013 was approximately 13 times greater than the incidence rate among U.S.-born persons, and the proportion of TB cases occurring in foreign-born persons continues

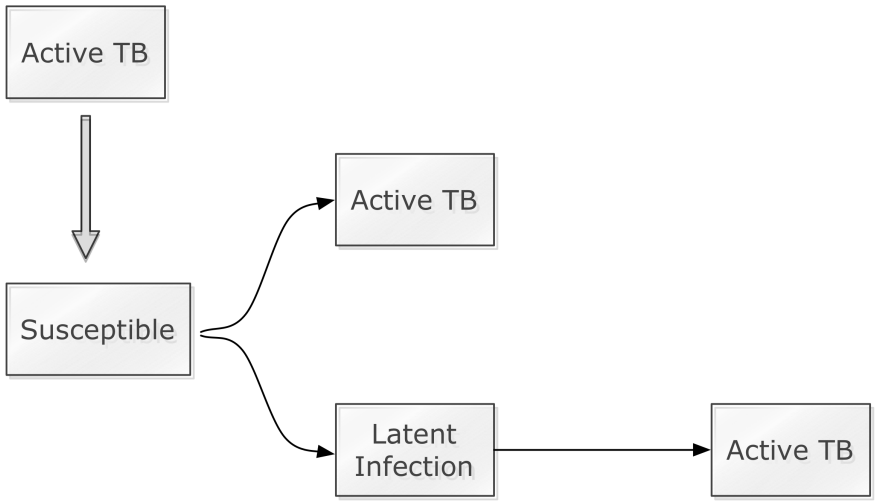


Figure 1.1: The dynamics of the transmission of TB disease. The individual without any TB infection is denoted as Susceptible. Only a small fraction of people with the initial infection of the MTBC will develop active TB disease. The remainder will have the latent TB infection and there is a 5% risk to progress to active TB through endogenous reactivation.

to increase, reaching 64.6% in 2013 [10]. There were 6,274 active TB cases of foreign-born persons reported in 2012 in the U.S., among those 21% are from Mexico, 12% from the Philippines and 9% from India [11]. As of 2012, among the foreign born TB patients in the United States, the top 7 leading countries of birth are shown in Figure 1.2 [11]. Since immigrants post the most significant burden to the TB, having an effective system of TB surveillance among immigrants in the United States will be an urgent task.

1.3 Genetic Diversity

As a development of biotechnology, DNA genotyping has became a crucial tool in TB research. The MTBC has great genetic diversity. Strains of MTBC are associated with different geographic regions. Research has shown that the different strains of MTBC display distinct levels of virulence and drug resistance, which can be translated into importance phenotypic differences [8, 12]. For example, experiments on guinea pigs have shown that the low-virulence strains from India are also less

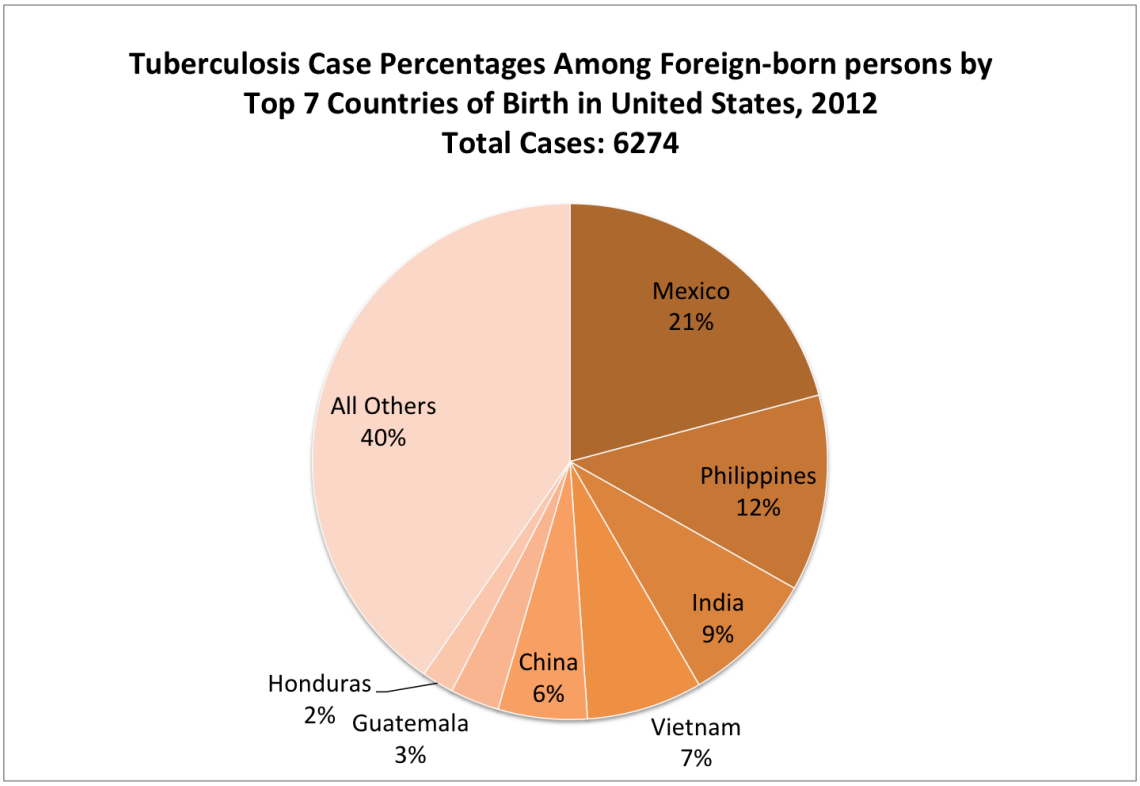


Figure 1.2: The top 7 leading countries of birth in the reported TB cases in the United State 2012.

infectious [12].

The genetic diversity of MTBC not only provides indicative information for the phenotype, but also offers evidence for transmission from person to person. Since the mutation rate of the MTBC is slow relative to the transmission rate [13], it can be assumed that the genetic information of the MTBC remain unchanged in the event of transmission. For instance, for two TB patients, we could hypothesize that one has transmitted the disease to the other. The fact that these two patients share the same MTBC genotype supports the hypothesis, while transmission is impossible if the two have different MTBC genotypes [14, 15]. Contact investigations are used to further investigate if transmission has occurred between two patients. Deciding when more extensive contact investigation are needed is a fundamental issue in TB control.

1.3.1 MTBC Genotyping

As discussed above, the discrimination between different genetically related MTBC strains plays a crucial role in monitoring TB disease spread. In order to identify different strains, we must access the genetic information of MTBC. The fundamental way to do so is to sequence the whole genome [16]. The MTBC was sequenced in 1998 [17]. However, sequencing the whole genome is both labor intensive and time consuming. Moreover, comparing the whole sequence data is computationally expensive. It is reasonable to only identify the genome loci which contain rich discriminating information for the purpose of tracking disease spread. DNA genotyping technology utilizes this idea to create DNA fingerprints. Here below, three popular and efficient MTBC genotyping techniques are presented.

RFLP: Restriction Fragment Length Polymorphism (RFLP) is the standard approach for genotyping MTBC, recognized as the "gold standard" [16]. The genome of MTBC often contains an insertion sequence named IS6110. RFLP analysis studies the distribution of the insertion sequence in different strains. The discrimination of the strains of MTBC is allowed based on the distributions. However, the RFLP analysis has some limitations. Despite having been standardized 10 years ago, the RFLP profiles are still difficult to compare across laboratories, which restricted the global application of the method [18]. Moreover, IS6110-based genotyping requires subculturing the isolates for several weeks to obtain sufficient DNA [19].

Spoligotyping: Spoligotyping is a PCR-based technique. It explores the polymorphism in the direct-repeat(DR) region to distinguish between strains [20]. The DR region contains 36-bp repeats which are separated by up to 43 non-repetitive 31-41 bp length sequences called spacers [20, 21]. Strains differ in terms of the presence and absence of specific spacers. The spoligotype of a strain is represented as a 43-bit long binary string, with a 0 representing absence and 1 representing presence of a spacer. A key fact about the mutation of spoligotypes is that once a spacer is lost, it is extremely unlikely to be regained. Therefore, it is hypothesized that spoligotypes evolve by deletion of a single or multiple contiguous spacers, whereas insertion is very unlikely [20–22]. Unfortunately, spoligotyping remains less discriminant than

IS6100- RFLP when used alone [23], yet it can be improved when using together with another method: MIRU-VNTR.

MIRU-VNTR: MIRU-VNTR typing is a variable number of tandem repeat analysis for bacterial typing scheme based on different classes of interspersed genetic elements named mycobacterial interspersed repetitive units (MIRUs). MIRU is a 46-100 bp DNA sequence dispersed within the intergenic regions of the MTBC genome as tandem repeats. MIRU-VNTR typing is based on the number of repeats observed at certain identified polymorphic loci. The degree of differentiation between strains depends on the number of loci used [20, 24]. A system of 12 loci (MIRU locus 2, 4, 10, 16, 20, 23, 24, 26, 27, 31, 39, 40) is currently the most common standard and has been integrated in TB control systems on a national scale [24].

We will use this 12 loci system in this thesis. Each MIRU result is reported as 12-character designations, each character corresponding to the number of repeats at one of 12 MIRU loci in the order listed previously. Each MIRU locus has possible repeat numbers from 0 to 14. We use the convention of representing the number of repeats which are greater than 10 as letters. Therefore, 0-14 will be represented as $\{0, 1, 2, ..., 9, a, b, c, d\}$. The exception is MIRU locus 4, which contains a variable number of 77-bp repeated unit followed by an invariable number of 53-bp units. In some isolates, the 53-bp units will be deleted. In order to differentiate the two cases, letters z to r are used to represent repeat number 1 to 9 for the repeats with the 53-bp units deleted respectively, e.g.,'z' is used for 1 repeat without the 53 - bp units, 'y' for 2, 'x' for 3 and so forth [25]. There is a total of 9 possible extra values represented by letters 'z' through 'r'. An example of a MIRU profile could be: 2x22323a2345.

MIRU-VNTR is easily reproducible and time efficient. Moreover, its stability allows the tracking of outbreak episodes, laboratory cross-contamination, or relapses [26]. When used in combination with spoligotypes, it provides a powerful method for strain discrimination.

1.4 Our Contributions

This thesis studies TB from two aspects. We first study the genome of the MTBC isolates. We take advantage of the massive database of the DNA fingerprints of MTBC to infer mutations among the isolates. Spoligotype is used to discover the mutation directions. Based on the mutation directions, we collected information on putative MIRU mutations. We find that how one repeat number changes in a mutation actually depends on the value of the repeat number. For example, large repeat numbers tend to decrease, while small values tend to increase in a mutation. We also found that, for 6 out of 14 repeat numbers, the most frequent change in a mutation is +1/-1, i.e. they either gain or lose a single repetitive unit. Contrary to the previous studies, which assume MIRU evolves identically across different loci [13], we find that the dynamics of the evolution for different loci are different. For instance, in a mutation, repeat numbers will change to a few certain values regardless what the initial values are. We call these "sticky" values. The "sticky" values are different for different loci. A discrete time Markov chain model is built to better understand the evolution if MIRU. The theoretical stationary distribution of the repeat numbers computed based on the Markov Chain model are compared with the sample distribution. This comparison gives an idea of the future variability to be expected in MIRU evolution by locus. The evolution rate is studied under the framework of Markov chain. We found that locus MIRU 24 is the most stable one with slowest evolution rate, while MIRU 27 is the least stable one with highest rate.

After looking into the genome of MTBC, we take a step back and investigate how TB spreads. Different than other mathematical models for TB epidemiology, which study the transmission dynamics at the population level [27], we build a model which studies the spread on an individual patient level. With the help of the DNA fingerprinting technology, TB patients can be clustered into small groups with size of 1-10. We define a cluster to be a group in which patients have the same spoligotype and RFLP. Doing this allows us to study TB spread in a more detailed manner. Our model utilizes the information we observe from the immigrant TB patients to estimate whether a specific patient in a cluster entered the country with or without latent TB infection. This will help TB control and surveillance. We built the model

in two steps: first: a detailed model with exact computation is built to allow us to understand the dynamics; second: a mean-field style approximation method is used to simplify the computation. In the simulation, the model is shown to have a good performance in identifying whether a foreign born patient is latently infected at entry or not. This model can help healthcare worker to identify the clusters where transmissions are most likely to occur, so that they can prioritize the clusters and allocate limited stuffs and resources for further investigation. The model is applied on the patient data collected from New York City in 2001-2007. Results show that the model successfully identified clusters with transmissions.

1.5 Organization

This thesis is organized in the following way. Chapter 2 covers our study on the MIRU evolution. Chapter 3 lays out the theoretical background on our model on the TB spread, including the exact model and the approximation model. The characteristics of the patient data posed some limitations of parameters estimation. This is discussed in Chapter 4. In Chapter 5, the TB spread model is tested on simulated data. The model is trying to identify a patient entering the country susceptible, indicating that the patient was infected by someone in the country, versus the patient entering with latent infection. The performance is measured by Receiver Operating Characteristic (ROC) curve. Also, the model is applied on the data from New York City. Finally, chapter 6 concludes research findings and points out the future research directions rooting from this study.

CHAPTER 2
MIRU Evolution

2.1 Introduction

In this study, we exploit massive databases of MTBC shared types characterized by both spoligotype and MIRU in order to examine how MIRU loci evolve. Our strategy is to use spoligotypes to determine evolutionary direction of potential evolutionary events. We look for isolate pairs with that have lost exactly one spacer and that have changed one or zero MIRU loci. We examine a joint data set of 14,453 isolates gathered from United States Centers for Disease Control (CDC) [28] and from Institute Pasteur SITVIT [29] to determine 41,604 of these potential pairs. The data from CDC are collected by the TB-Insight project (http://tbinsight.cs.rpi.edu) and it appeared in a previous study [28]. Then we performed two separate studies of the result. The first analysis examines the frequency and mutation variability in the number of repeats and how mutation depends on the number of repeats. The results suggest that the probability of mutation varies by the number of repeats. The second analysis examines the frequency and variability in mutation by loci. These results show that the variability of repeats differs by loci. In the second analysis the mutations in the repeat numbers of different MIRU loci are analyzed using the framework of discrete time Markov Chains. The theoretical stationary distribution of the repeat numbers computed based on the Markov Chain model are compared with the sample distribution to give an idea of the future variability in MIRU evolution by locus.

2.2 Method

We propose a model to utilize the information from both MIRU-VNTR typing and spoligotyping to infer MIRU mutations. Each MTBC isolate is genotyped by spoligotye by 43 binary digits capturing the absence and presence of spacers and one 12 loci MIRU profile, which is a 12 digit character string. We consider all possible pairs of isolates consisting of a parent and child isolate which satisfy the following

two rules.

As discussed above, given a mutation happened, the DR region of MTBC is likely to gain a spacer, while losing one is nearly impossible. This means only mutations from 1 to 0 are possible among spoligotypes. This fact is used to discover the mutational directions among different MTBC isolates. A mutation between two isolates is defined when the following two rules are satisfied.

Rule 1. *The* parent *isolate and* child *isolate have 42 identical spacers and one spacer that is lost, i.e. the parent has a 1 for that spacer and the child has a 0 in the changed position of the shared spoligotype.*

Rule 2. *The parent and child have 11 identical MIRU locus alleles and on one locus changes by at least one spacer. The evolutionary direction is infered to be from the parent to the child as determined by the spoligotype rule.*

Figure 2.1 shows an example of one incidence of the mutation.

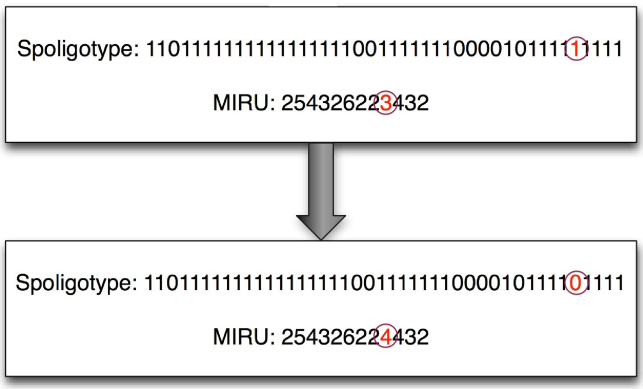


Figure 2.1: An example of one mutation. Where the 39th spoligotype (in circle) goes from 1 to 0 and the 9th MIRU goes from 3 to 4 (in circle). Note that we only allow one change in both numbers

Further understanding can be gained by examining the distribution of child values per each locus. The change from parent repeat number i to child number j were calculated. Parent-child events were collected from the data set for each locus. The numbers of transitions from i to j repeats were counted for every $i, j \in S$. For each MIRU locus k, a matrix $M^{(k)}$ is built based on the counts.

$$M^{(k)} = \begin{bmatrix} u_{00}^{(k)} & u_{01}^{(k)} & u_{02}^{(k)} & ... & u_{0N}^{(k)} \\ u_{10}^{(k)} & u_{13}^{(k)} & u_{12}^{(k)} & ... & u_{1N}^{(k)} \\ u_{20}^{(k)} & u_{21}^{(k)} & u_{22}^{(k)} & ... & u_{2N}^{(k)} \\ u_{30}^{(k)} & u_{31}^{(k)} & u_{32}^{(k)} & ... & u_{3N}^{(k)} \\ . & & & & \\ . & & & & \\ u_{N0}^{(k)} & u_{N1}^{(k)} & u_{N2}^{(k)} & ... & u_{NN}^{(k)} \end{bmatrix} \tag{2.1}$$

where each entry $u_{ij}^{(k)}$ represents the number of counts that i turned into j repeats at locus k. For example, $u_{23}^{(4)} = 130$ means that there are 130 times of repeat number 2 turning into 3 at MIRU locus 4. We will refer $M^{(k)}$ as the transition counts matrix for MIRU locus k.

2.2.1 Markov Chains

The model in this study is based on a discrete time Markov Chain, which is a memoryless stochastic process [30]. We will first introduce some of the important properties of Markov Chain here.

Definition: A stochastic process is the evolution of some random variables over time. Consider a sequence of random variables X_n, $n = 0, 1, 2, ...$, with n representing the time, which is called the *epoch* of the process. X_n takes values in a finite set $S = \{0, 1, 2, ..., N\}$, which is called the *states* of the process. The transition probability is the conditional probability of $X_n = i_n$ given the first $n - 1$ epochs.

$$P(X_n = i_n | X_0 = i_0, X_1 = i_1, ..., X_{n-1} = i_{n-1}) \tag{2.2}$$

A Markov Chain is a stochastic process with the Markov property, which is that the distribution of X_n only depends on the previous epoch.

$$P(X_n = i_n | X_0 = i_0, X_1 = i_1, ..., X_{n-1} = i_{n-1}) = P(X_n = i_n | X_{n-1} = i_{n-1}) \tag{2.3}$$

This is also called the *memoryless property* and the expression on the right hand side is called a *transition probability*. A *time-homogeneous Markov Chain* is a process such that the transition probability is constant at different epochs.

$$P(X_n = j|X_{n-1} = i) = p(i, j) \tag{2.4}$$

We will assume time-homogeneity in our MIRU model. A *transition probability matrix* is a $(N+1)\times(N+1)$ matrix P, with entries equal to the transition probabilities, i.e. $p_{ij} = p(X_n = j|X_{n-1} = i)$.

Communication Class: The initial distribution of the states, $\boldsymbol{q}_0$, is a vector with entries $q_0(i) = P(X_0 = i)$. The distribution of states at the next epoch will be $\boldsymbol{q}_1 = P'\boldsymbol{q}_0$, where P' is the transpose of the transition probability matrix P. Moreover, the probability distribution for the state at the n^{th} epoch will be $\boldsymbol{q}_n = P'^n\boldsymbol{q}_0$.

Two states i and j are said to *communicate* with each other, if there exist $m, n > 0$, such that $p^m_{ij} > 0$ and $p^n_{ji} > 0$, where $p^n_{i,j}$ is the (i, j) entry of P^n. That is equivalent saying if i and j communicate, there is a path to reach j starting from i and vice versa.

A *communication class* is a set of states such that any pair of the states communicate with each other. A Markov Chain with only one communication class is called *irreducible*. A Markov Chain could have multiple communication classes. There are two types of communication classes: *recurrent* and *transient*. If the chain starts from a recurrent class, then it stays in this class forever. On the other hand, if the chain starts in the transient class, it will have probability one to leave it in some future epoch.

Periodicity: The period of a state i of a Markov chain is defined as the following,

$$d = gcd\left\{n : X_n = i|X_0 = i\right\} \tag{2.5}$$

where *gcd* stands for greatest common divisor. A state with period of 1 is said to

be aperiodic. A communication class is aperiodic if all of its states are aperiodic. A communication class only needs one aperiodic state to imply all the states are aperiodic [30].

Long-Range Behavior: Let P be the transition probability matrix of an irreducible Markov Chain or recurrent communication class and $\boldsymbol{q}_0$ be its initial distribution. We have the following

$$\boldsymbol{q}_n = P'^n \boldsymbol{q}_0 \tag{2.6}$$

If the Markov Chain (recurrent class) is aperiodic and irreducible, then there exists a unique probability distribution such that

$$\boldsymbol{\pi} = \lim_{n\to\infty} \boldsymbol{q}_n = \lim_{n\to\infty} P'^n \boldsymbol{q}_0 \tag{2.7}$$

Such $\boldsymbol{\pi}$ is called the *stationary distribution* of the Markov Chain (recurrent class). There are two important properties about the transition probability matrix P [31],

Lemma 1. *Provided the Markov Chain is irreducible and aperiodic, the corresponding probability transition matrix will have the following properties:*

- *Exactly one of the eigenvalues of P has the value of 1.*
- *All the eigenvalues of P have absolute values less than 1.*

MIRU Markov Chain Model: Recall that we have the transition counts matrix for MIRU locus k, $M^{(k)}$, which is defined as equation (2.1). The standard procedure for fitting a Markov chain model to a data set is to estimate the probability transition matrix, $\hat{P}^{(k)}$ from the transition count matrix, i.e. $M^{(k)}$. Let $\hat{p}_{ij}^{(k)}$ be the entry at row i and column j of $\hat{P}^{(k)}$. Viewing each row of $M^{(k)}$ as a sample of the multinomial distribution with parameters $[p_{i0}^{(k)}, p_{i1}^{(k)}, ..., p_{iN}^{(k)}]$, $\hat{p}_{ij}^{(k)}$ will be an estimation for $p_{ij}^{(k)}$ and the corresponding standard error will be $\hat{\sigma}_{ij}^{(k)}$. Here we use Laplace smoothing on the transition count matrix, which is equivalent to add 1 to each $u_{ij}^{(k)}$ [32]. $\hat{P}^{(k)}$ and $\hat{\sigma}_{ij}^{(k)}$ are defined as follows [33].

$$\hat{p}_{ij}^{(k)} = \frac{u_{ij}^{(k)} + 1}{\sum_{j \in S} u_{ij}^{(k)} + N + 1} \tag{2.8}$$

$$\hat{\sigma}_{ij}^{(k)} = \sqrt{\frac{\hat{p}_{ij}^{(k)}(1 - \hat{p}_{ij}^{(k)})}{\sum_{j \in S} u_{ij}^{(k)} + N}} \tag{2.9}$$

where $N+1$ is the dimension of the state space is (e.g. the possible repeat numbers are $0, 1, 2, ..., N$), therefore it appears in the denominator of equation (2.8). The insignificant estimations, which means $\hat{p}_{ij}^{(k)} < 2\hat{\sigma}_{ij}^{(k)}$, are set to 0.

Let $\boldsymbol{q}_m^{(k)} = [q_{1,m}^{(k)}, q_{2,m}^{(k)}, ..., q_{n,m}^{(k)}]'$ be the column vector representing the distribution of the repeat numbers at the m^{th} generation at locus k. The stationary distributions of the communication classes of every MIRU locus can be obtained by the following equation [30]:

$$\boldsymbol{\pi}^{(k)} = \mathbf{1}(I - \hat{P}^{(k)} + ONE)^{-1} \tag{2.10}$$

where $\mathbf{1}$ is a vector of ones, I is the identity matrix, $\hat{P}^{(k)}$ is the transition probability matrix of the major communication class and ONE is a matrix with all ones.

Error Analysis: Note that we estimated the transition probability matrix $P^{(k)}$ based on the transition counts $M^{(k)}$. The error from this estimation is propagated to the computation of the stationary distribution $\boldsymbol{\pi}^{(k)}$. A Monte Carlo simulation scheme is used to evaluate this error. We are going to use Dirichlet distribution to sample. It is a distribution on multivariate random variable $\mathbf{X} = [x_1, x_2, ..., x_n]$, $x_i > 0$ and $\sum_{i=1}^{n} x_i = 1$ and parameter $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_n]$, $\alpha_i > 0$. The probability density function is

$$\begin{aligned} \mathrm{Dir}(\mathbf{X}|\boldsymbol{\alpha}) &= \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{n} x_i^{\alpha_i - 1} \\ B(\boldsymbol{\alpha}) &= \frac{\prod_{i=1}^{n} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{n} \alpha_i)} \end{aligned} \tag{2.11}$$

where $\Gamma(x)$ is the gamma function, $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$.

For each $i \in S$, a row vector $\tilde{\boldsymbol{p}}_i^{(k)} = [\tilde{p}_{i0}^{(k)}, \tilde{p}_{i1}^{(k)}, ..., \tilde{p}_{iN}^{(k)}]$ is simulated with the

Dirichlet distribution $\text{Dir}(\boldsymbol{X}|u_{i0}^{(k)}, u_{i1}^{(k)}, ..., u_{iN}^{(k)})$, where $\boldsymbol{X} = [x_0, x_1, ..., x_N]$ is a $N+1$-dimensional random variable. For each sample from the Dirichlet distribution $\tilde{\boldsymbol{p}}_i^{(k)} = \boldsymbol{X}$. There are a few reasons why we use the Dirichlet distribution for sampling,

- A Dirichlet random variable $\boldsymbol{X}$ will be a categorical probability. This means that $0 \leq x_i \leq 1$ for $i = 1, 2, ..., d$ and $\sum_{i=1}^{d} x_i = 1$, where x_i is the i^{th} element of the d-dimensional random variable $\boldsymbol{X}$.

- Let $\tilde{\boldsymbol{p}}_i^{(k)} = [\tilde{p}_{i0}^{(k)}, \tilde{p}_{i1}^{(k)}, ..., \tilde{p}_{iN}^{(k)}]$ be a sample from $\text{Dir}(\boldsymbol{X}|u_{i0}^{(k)}, u_{i1}^{(k)}, ..., u_{iN}^{(k)})$. The expectation of $\tilde{p}_{ij}^{(k)}$ will match that of $\hat{p}_{ij}^{(k)}$ and the standard deviation will match approximately, as shown in the following

$$E[\tilde{p}_{ij}^{(k)}] = \frac{u_{ij}^{(k)}}{\sum_{j \in S} u_{ij}^{(k)}} \tag{2.12}$$

$$\sqrt{Var[\tilde{p}_{ij}^{(k)}]} = \sqrt{\frac{\hat{p}_{ij}^{(k)}(1 - \hat{p}_{ij}^{(k)})}{\sum_{j \in S} u_{ij}^{(k)} + 1}} \tag{2.13}$$

Note that the standard deviation of the simulated $\tilde{p}_{ij}^{(k)}$ and the standard error of the estimated $\hat{p}_{ij}^{(k)}$ (the results of equation (2.9) and (2.13)) are different. However, due to the large values of $\sum_{j \in S} n_{ij}^{(k)}$, the differences are negligible.

The matrix $\tilde{P}^{(k)}$ with entries $\tilde{p}_{ij}^{(k)}$ will have the expectation of $\hat{P}^{(k)}$. The standard deviations of each entry of $\tilde{P}_{ij}^{(k)}$ is approximately equal to the standard error of the MLE estimation. For each $\hat{P}^{(k)}$, $\tilde{P}^{(k)}$ is simulated 10000 times and a stationary distribution $\tilde{\boldsymbol{\pi}}^{(k)}$ is computed using the same equation (2.10). The confidence intervals of $\boldsymbol{\pi}^{(k)}$ are constructed based on the standard deviations of the computed $\tilde{\boldsymbol{\pi}}^{(k)}$.

Forward Simulation: Given the sample distribution of the repeat number of a MIRU locus k is $\boldsymbol{q}_0^{(k)}$, as discussed in the previous section, the distribution of the n^{th} generation will be $\boldsymbol{q}_n^{(k)} = \hat{P}'^n \boldsymbol{q}_0^{(k)}$. $\boldsymbol{q}_n^{(k)} \to \boldsymbol{\pi}^{(k)}$ as $n \to \infty$, where $\boldsymbol{\pi}^{(k)}$ is the stationary distribution computed in equation (2.10). The rates of convergence indicate how fast

each loci approaches to the stationary state. Let $\hat{P}^{(k)}$ be the maximum likelihood estimation of the transition probability matrix of an aperiodic, irreducible Markov chain. Suppose $\hat{P}'^{(k)}$ has K eigenvalues, $|\lambda_1| > |\lambda_2| \geq ... \geq |\lambda_K|$, and by Lemma 1, $\lambda_1 = 1$. For any given initial distribution $\boldsymbol{q}_0^{(k)}$, we have the following,

$$\begin{aligned} \boldsymbol{q}_n^{(k)} &= \hat{P}'^{(k)n}\boldsymbol{q}_0^{(k)} \\ &= \lambda_1^n c_1 \boldsymbol{v}_1 + \lambda_2^n c_2 \boldsymbol{v}_2 + ... + \lambda_K^n c_K \boldsymbol{v}_K \\ &\approx \lambda_1^n c_1 \boldsymbol{v}_1 + \lambda_2^n c_2 \boldsymbol{v}_2 + o(\lambda_2^n c_2 \boldsymbol{v}_2) \end{aligned} \tag{2.14}$$

where $\boldsymbol{v}_i$ is the eigenvector of $\hat{P}'^{(k)}$ corresponding to λ_i and $||\boldsymbol{v}_i|| = 1$. Note that $\lambda_1 = 1$ and $|\lambda_2| < 1$, so when $n \to \infty$, $\boldsymbol{q}_n^{(k)} \to c_1\boldsymbol{v}_1 = \boldsymbol{\pi}^{(k)}$ and $\lambda_2^n c_2 \boldsymbol{v}_2 \to 0$. Let $r = -log(|\lambda_2|)$, then $|\lambda_2|^n = e^{-rn}$. This is an exponential decay function in terms of the number of generation n. Therefore, the rates of convergence can be measured by $-log(|\lambda_2|)$. If the distance between $\boldsymbol{q}_n^{(k)}$ and $\boldsymbol{\pi}^{(k)}$ is measured by the 2-norm, $\|\boldsymbol{q}_n^{(k)} - \boldsymbol{\pi}^{(k)}\|$ we have the following,

$$\|\boldsymbol{q}_n^{(k)} - \boldsymbol{\pi}^{(k)}\| = \|\lambda_2^n c_2 \boldsymbol{v}_2 + o(\lambda_2^n c_2 \boldsymbol{v}_2)\| \tag{2.15}$$

Let $\boldsymbol{x} = [x_1, x_2, ..., x_n]$ be a vector with n entries. The 2-norm of $\boldsymbol{x}$ is defined as follows,

$$||\boldsymbol{x}|| = \sqrt{\sum_{i=1}^{n}(x_i)^2} \tag{2.16}$$

We could compute the number of steps, starting from the initial distribution $\boldsymbol{q}_0^{(k)}$, needed for the distance to decrease to the magnitude of certain threshold value ϵ. This can be done by simply solving $|\lambda_2^n c_2| \leq \epsilon$. We have,

$$n \geq \frac{\ln(\epsilon) - \ln(|c_2|)}{\ln(|\lambda_2|)} \tag{2.17}$$

where n is the minimum of steps for the distance between the current and stationary distributions to drop within the magnitude of ϵ.

However, the 2-norm ($\|\cdot\|$) is not the standard way to measure the distance between distributions. Instead, it is usually measured by Kullback-Leibler (KL) divergence [34]. The KL-distance between $\boldsymbol{q}_n^{(k)}$ and $\boldsymbol{\pi}^{(k)}$ is defined as the following,

$$D(\boldsymbol{q}_n^{(k)}|\boldsymbol{\pi}^{(k)}) = \sum_{i=1}^{d} \ln\left(\frac{\varphi_{n,i}^{(k)}}{\psi_i^{(k)}}\right)\varphi_{n,i}^{(k)} \tag{2.18}$$

where $\varphi_{n,i}^{(k)}$, $\psi_i^{(k)}$ are the i^{th} elements of $\boldsymbol{q}_n^{(k)}$ and $\boldsymbol{\pi}^{(k)}$ respectively and d is the dimension of two distributions. Using the results from equation (2.14) and taking the first two terms of the *Taylor Expansion* of the logarithm, we can put $D(\boldsymbol{q}_n^{(k)}|\boldsymbol{\pi}^{(k)})$ in the following form:

$$\begin{aligned} D(\boldsymbol{q}_n^{(k)}|\boldsymbol{\pi}^{(k)}) &= \sum_{i=1}^{d} \ln\left(\frac{\psi_i^{(k)} + \lambda_2^n c_2 \omega_2^i + o(\lambda_2^2)}{\psi_i^{(k)}}\right)\varphi_{n,i}^{(k)} \\ &\approx \sum_{i=1}^{d} \ln\left(1 + \frac{\lambda_2^n c_2 \omega_2^i}{\psi_i^{(k)}}\right)\psi_i^{(k)} \\ &\approx \sum_{i=1}^{d} \left[\frac{\lambda_2^n c_2 \omega_2^i}{\psi_i^{(k)}} - \frac{1}{2}\left(\frac{\lambda_2^n c_2 \omega_2^i}{\psi_i^{(k)}}\right)^2\right]\psi_i^{(k)} \\ &= \lambda_2^n c_2 \sum_{i=1}^{d} \omega_2^i + o(\lambda_2^n) \end{aligned} \tag{2.19}$$

where ω_2^i is the i^{th} element of $\boldsymbol{v}_2$. We can compute the number of steps for $D(\boldsymbol{q}_n^{(k)}|\boldsymbol{\pi}^{(k)})$ to decrease to within the magnitude of ϵ by solving $|\lambda_2^n c_2 \sum_{i=1}^{d} \omega_2^i| \leq \epsilon$. This will give us

$$n \geq \frac{\ln(\epsilon) - \ln(|c_2|) - \ln(|\sum_{i=1}^{d} \omega_2^i|)}{\ln(|\lambda_2|)} \tag{2.20}$$

In addition to these analytically computed step numbers n, we will also perform the forward simulation. The values $D(\boldsymbol{q}_n^{(k)}|\boldsymbol{\pi}^{(k)})$ will converge to 0 as $\boldsymbol{q}_n^{(k)} \to \boldsymbol{\pi}^{(k)}$. The values of $D(\boldsymbol{q}_n^{(k)}|\boldsymbol{\pi}^{(k)})$ are computed for every n. We count the number of

steps, $\hat{n}$, it takes for $D(\boldsymbol{q}_n^{(k)}|\boldsymbol{\pi}^{(k)})$ drop below ϵ (if $\lambda_2 > 0$).

2.3 Results

To infer MIRU evolution tendencies, we analyzed a collection of 14,453 isolates geneotyped by spoligotype and MIRU provided by the United States Centers for Disease Control from a collection of TB isolates from the patients in the United States collected from 2004 to 2007 [35] and by Institute Pasteur de Guadeloupe from the SITVITWEB collection [36]. A total of 41,604 pairs of parent-child was found based on the two rules we defined, **Rules** (1) and (2).

2.3.1 Analysis by Repeat Number

The distribution of the repeat numbers in our data set is shown in Figure 2.2. Repeat number 1-7 appear most frequently. These 7 numbers take 99.27% of the total numbers. The mode is 2, which accounts for 35.89% of the total number of appearances.

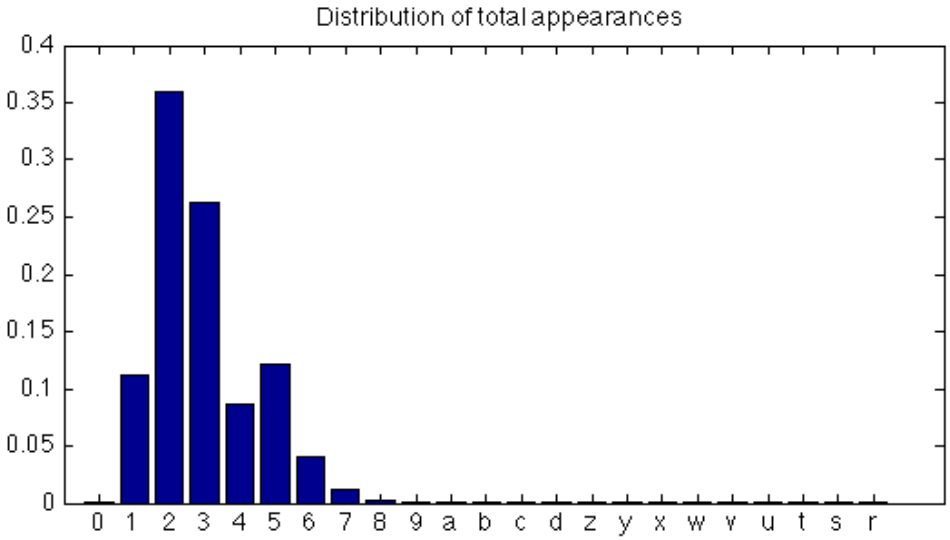


Figure 2.2: The distribution of the appearance of repeat numbers in the data. Repeat 1- 7 are the most frequent ones, while 2 is the mode. 1-7 make 99.27% of the total appearances, while 2 makes up 35.89%.

Figure 2.3 (top) shows the distribution of the repeat numbers observed in the

parents. The majority of the repeat numbers are from 1,2,3,4,5,6,7 with the mode at 2. The distribution of the appearance of the repeat numbers in the parents closely corresponds to the distributions of repeat numbers observed in the data set since only one locus changes from parent to child per mutation event and all loci of the parent are counted. The distribution of repeat numbers of parent alleles that are inferred to mutate is quite different. Figure 2.3 (bottom) examines the probability that a repeat number changes from parent to child in one of the inferred mutation events. Repeat numbers 1 through 6 are observed to mutate less than 17% of the time. Repeat 7 changes 35.19%, repeat 'a' changes 40% and 0,8,9,b,c,d all have changing percentage higher than 74%. This result agrees with prior study which observed that large repeat numbers have higher changing probability [13]. Note that in our finding, repeat number 0 is an exception. Although it is the smallest repeat number, it has a 89.54% chance to change in the event of mutation.

We also observed that in a mutation, the distribution of the child's repeat number depends on the parents'. Figure 2.4 shows the distribution of child's repeat numbers for each of the possible parent repeat numbers given that a change of at least one occurs. We exclude the mutations when the repeat number from 0-d change to 'z'-'r', since these cases are rare: given the repeat number changes, 1.22% of 0 and 1.11% of 9 change to 'z'-'r', while others have a chance less than 1%. For repeat number 1,2,3,4,5 and 6, the most frequent change is +1/-1, i.e. 3 changes to 2 or 4. Repeat numbers 0-3 are more likely to increase than decrease, while 4 and greater tend to decrease. The probability to increase for repeat 3,4 and 5 are: 3: 54.88%, 4: 38.85%, 5: 26.72%. The complete increase/decrease probabilities for all the repeat numbers of the parents are shown in Table 2.1.

Based on our findings, in the event of mutation, small repeat numbers have a higher probability to increase while larger repeat numbers tends to decrease. This contradicts the prior study which assumes increase/decrease probabilities are equally likely for all repeat numbers [13].

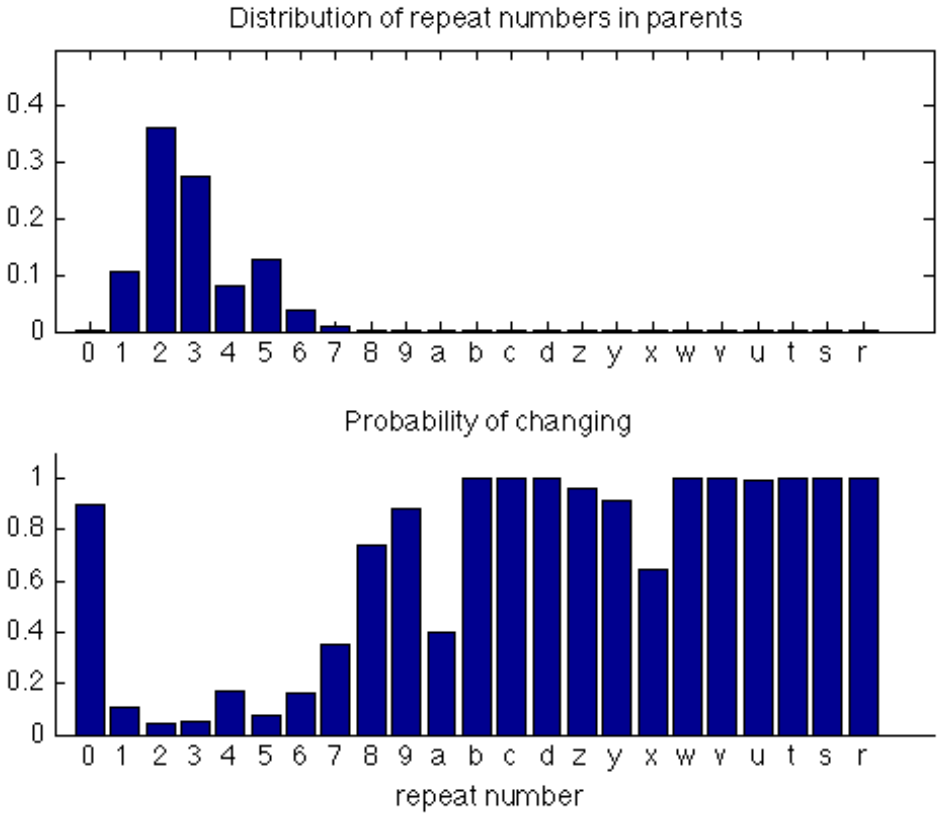


Figure 2.3: The top plot shows the distribution of the repeat number across all loci in the parents. The bottom plot shows the frequency with which a given repeat number change values from parent to child.

2.3.2 Analysis by Locus

In this section we investigate the difference in inferred MIRU mutations by locus. For each locus, we estimate the probability that the child isolate will have a change in that locus. As shown in Figure 2.5, MIRU 24 seems to be the most stable locus. In total of 41,604 mutations, only 0.58% of the repeat numbers at MIRU 24 changed their values. Loci 2, 20, 27 and 39 also have low change probability (less than 4.7%). Loci r, 4,16,23 and 31 have moderate rates of changes. (between 9.05% and 7.65%). On the other hand, MIRU 10, 26, 40 have a relatively high probability of changing repeat numbers (between 12.90% and 16.70%).

We collected the transition count matrix $M^{(k)}$ from the 41,406 pairs of parent-child data. Let $\hat{P}^{(k)}$ be the matrix resulting from normalizing the rows of $M^{(k)}$ with

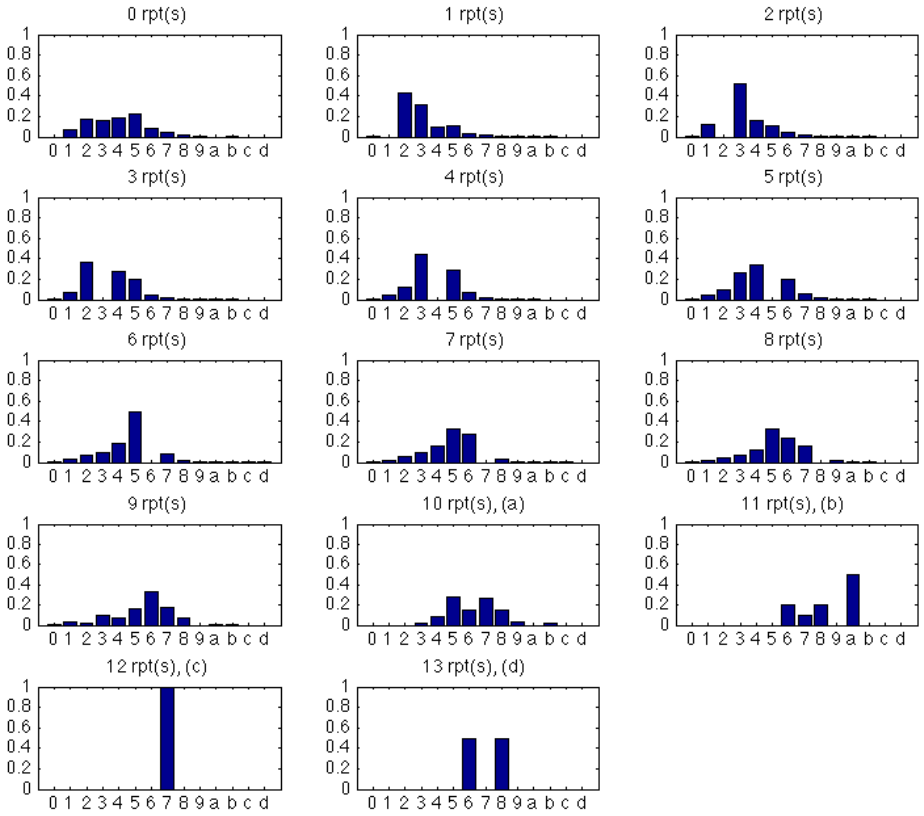


Figure 2.4: For each repeat number in the parents, given it changes in a mutation, the distribution of the child's repeat number. The mutations from 0-'d' to 'z'-'r' are rare (around 1%) and excluded.

Laplace smoothing [32], i.e. $\hat{p}_{ij}^{(k)} = \dfrac{u_{ij}^{(k)} + 1}{\sum_{j=0}^{N} u_{ij}^{(k)} + N + 1}$, where $u_{ij}^{(k)}$ is the number of counts of repeat number i changing to j at locus k. For example, an entry of $\hat{P}^{(24)}$: $p_{23}^{(4)} = 0.041$ means that at MIRU 4, given the current repeat number is 2, there is a 4.1% chance it will turn into 3 in one mutation. $\hat{p}_{ij}^{(k)}$ is an estimation to the true transition probability based on the data. To avoid noise, we want to only work with the significant estimations. With the standard deviation of $\hat{p}_{ij}^{(k)}$, $\hat{\sigma}_{ij}^{(k)}$, defined as in equation (2.9), we define an estimation of $\hat{p}_{ij}^{(k)}$ to be insignificant if $\hat{p}_{ij}^{(k)} < 2\hat{\sigma}_{ij}^{(k)}$. These insignificant estimations are set to 0. The matrix $\hat{P}^{(k)}$ is then used to define a discrete-time Markov chain $\{X_n\}_{n=1,2,\ldots}$, where $X_n^{(k)}$ models the k^{th} locus at the

Table 2.1: In the event of a mutation, the increase/decrease probabilities for each repeat number of the parent. Parents with small repeat number (0-3) have a higher probability to increase while parents with larger repeat number (4 or greater) tend to decrease.

Repeat No.	Decrease	Increase	Repeat No.	Decrease	Increase
0	0.0000	1.0000	7	0.9479	0.0521
1	0.0027	0.9973	8	0.9717	0.0283
2	0.1322	0.8678	9	0.9813	0.0187
3	0.4512	0.5488	10 (a)	0.9833	0.0167
4	0.6115	0.3885	11 (b)	1.0000	0.0000
5	0.7328	0.2672	12 (c)	1.0000	0.0000
6	0.8853	0.1147	13 (d)	1.0000	0.0000

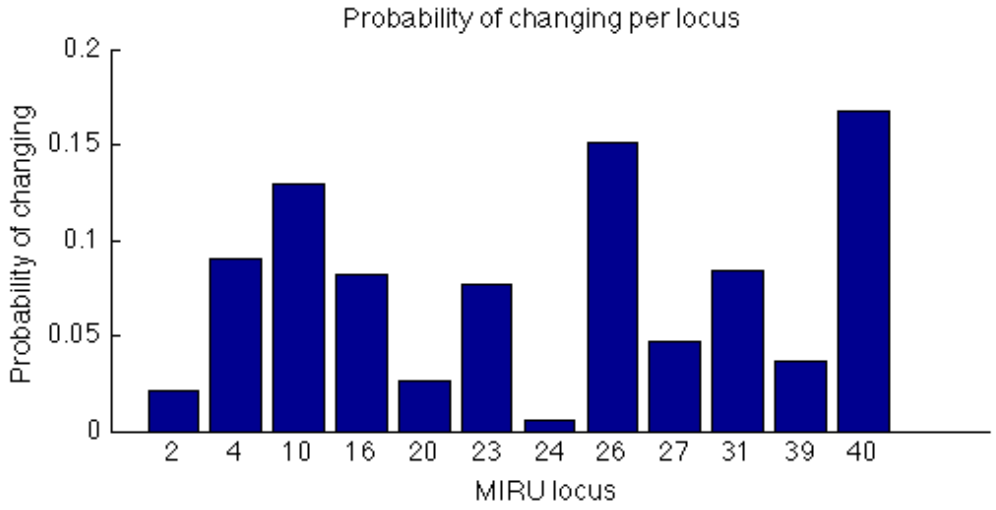


Figure 2.5: For each locus, the probability that the child's repeat numbers are different from the ones of the parent. MIRU 24 is the most stable locus as only 0.58% of repeat numbers change from parent to child. MIRU 40 is the least stable one, 16.7% of repeat numbers change in a mutation.

n^{th} generation. We define the Markov chain dynamics so that the probability of $X_{n+1} = j$ given $X_n = i$ is $\hat{p}_{ij}^{(k)}$. The $\hat{P}^{(k)}$ is called the *transition probability matrix* of the Markov Chain $\{X_n\}_{n=1,2,\ldots}$.

The heat-maps of transition matrices of each locus are shown in Figure 2.6 for all loci except MIRU 4 which is show in Figure 2.7. The heat-maps allows one to quickly visually assess differences between loci. Black indicates probabilities of

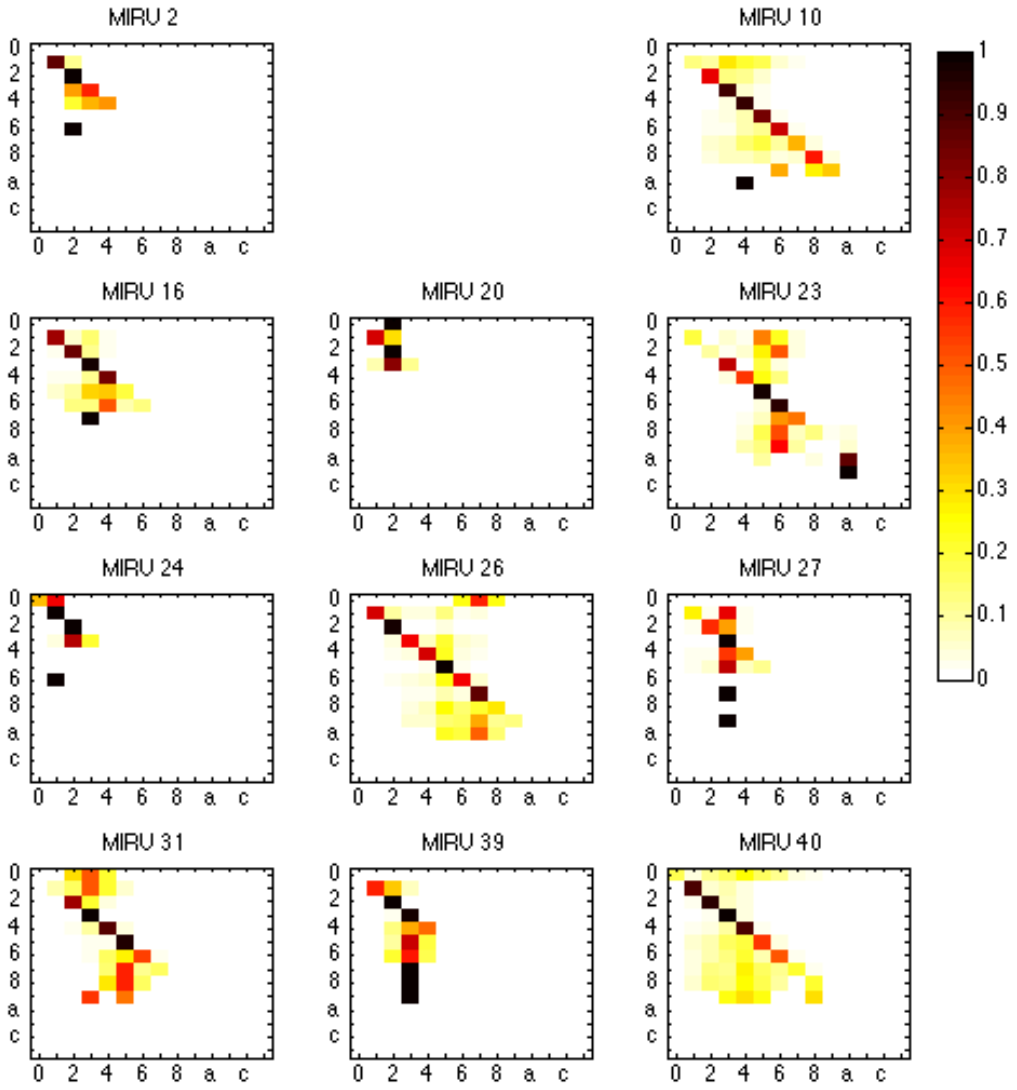


Figure 2.6: The heatmaps of the 11 transition probability matrices by loci (locus 4 is shown in a separate figure), where darker color means higher probabilities. Note that row number is the originating state and column number is the destination state under a mutation.

1 and white indicates probabilities of 0. The rows are the parent repeat numbers and the columns are the child repeat numbers. Dark values on the diagonal indicate repeats that tend to be stay at their current values; entries below the diagonal

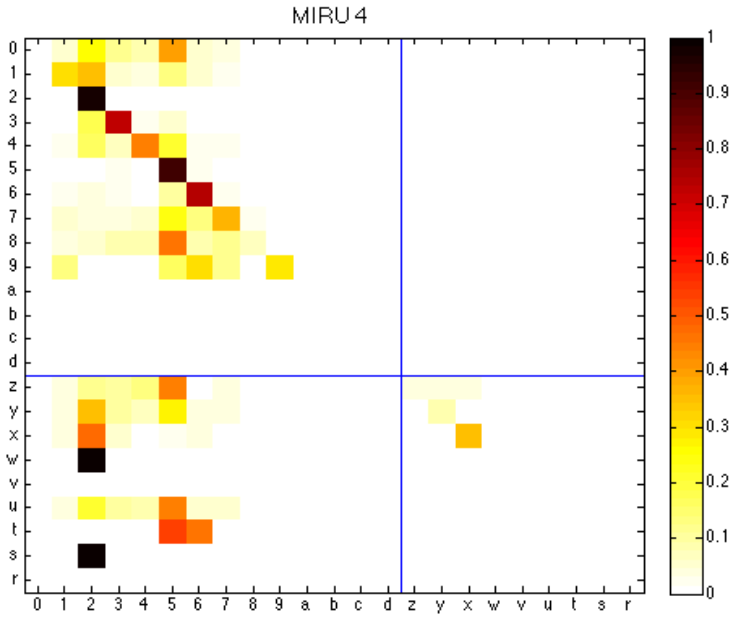


Figure 2.7: The heatmap of the transition probability matrix for MIRU locus 4. The lines separate the normal repeated units (0 - d) and the ones with 53-bp deleted (z - r).

indicate decreases and entries above the diagonal indicate increases. Entries one off the diagonal indicate values that change by one. Entries far from the diagonal indicate large changes. We offer several observations.

1. The heat-maps of MIRU 20 and 24 indicate that the observed values and transitions are confined to 0 to 4 repeats. Based on the inferred distribution, observed repeat values will remain small and mutations to and observations of large values (greater than 4) are anticipated to be very rare in the future based on our results. Also from our prior analysis in Figure 4, MIRU 20 and 24 are very stable in the sense that they are unlikely to change and when they do change they are likely remain at small numbers of repeats.

2. MIRU locus 24 is known to correspond to the TbD1 deletion. TbD1 is a DNA region that is present in ancestral strains (Indo-Oceanic, *M. bovis* and *M. Africanum*) but absent in modern strains (Euro-American, East Asian, and East African Indian) [37]. The modern strains will have less than 2 repeats

at MIRU 24 whereas ancestral strains more than 2. Although it is not visible from the heatmap, there are 98 pairs of parent-child relations, where MIRU 24 changes from 1 to 2, indicating a mutation from Modern to Ancestral strains, which is not expected. The lineage of both the parents and children are the same in these 98 mutations. Among the pairs, 1 is East-Asian, 3 are *Mycobacterium*- africanum, 6 are *Mycobacterium*-bovis, 7 are Euro-American and 81 are Indo-Oceanic.

3. MIRU 4, 10, 23, and 40 show a more complex structure with many off diagonal probabilities indicating a large number of mutations of different sizes and directions to a variety of children repeat values.

4. MIRU 4 extra repeats 'z' - 'r'. Recall that these characters represent the repetitive units without the 53-bp block. Our study shows that the mutations from 0 - 'd' to 'z'-'r' are rare. For repeats from 0 - 'd', the percentages of mutations to 'z'-'r' are the following, 9: 4.05%, 0: 3.42%, 7: 2.88%, all others have the percentages less than 1.5%. This results correspond to the heatmap shown in Figure 2.7, in which the upper right block is near empty. On the other hand, for the parents with repeat number 'z' - 'r', they will most likely change to values in 0 - 'd'. The percentages of values in 'z' - 'r' changing to 0 - 'd' are the following: 'x': 63.85%, 'y': 89.48%, 'z': 89.62%, all others have the percentages greater than 90%. This corresponds to the lower left block in Figure 2.7.

5. Most of the heatmaps display a common pattern: 1) they have darker diagonal entries, which represents the repeat number stays the same in a mutation; 2) they have one or more darker columns, which indicates some values are more common than others in the children. We name these values as "sticky values" and they will be discussed in the following section.

Sticky Values: Repeat numbers are more likely to mutate to certain values than others. We can see this in Figure 2.6. Colored columns indicates that certain child repeat numbers occur more frequently from a variety of values. For example, at

locus 27, if a repeat number is going to mutate, it is most likely for it to change to 3 regardless of the parent repeat value. This phenomenon is also observed for one or more repeat values at other loci. The values which other repeat numbers tend to mutate to are referred as "sticky" values in this study. We measure the stickyness of each repeat number by summing up the columns of the count matrix for each MIRU locus, $N^{(k)}$, while excluding the diagonal entries. The values are then normalized by sum of all these sums. For example the "stickyness" of each repeat number for MIRU locus 27 will be computed as the following:

1. Compute the columns sums of $N^{(27)}$ without the diagonal entries: $\text{Sum}_i = \sum_{j \neq i} N_{ij}^{(27)}$
2. Normalize the values by $\sum_{i=0}^{N} \text{Sum}_i$: $\text{Stickyness}_i = \frac{\text{Sum}_i}{\sum_{i=0}^{N} \text{Sum}_i}$

The results are shown in Figure 2.8. MIRU 2 ,20 and 24 have a sticky value at 2; MIRU 16, 27, 31, 40 at 3; MIRU 10 at 4, MIRU 26 at 5. Moreover, three loci have two sticky values: MIRU 4(2,5), MIRU 23(5,6), MIRU 39(2,3)

Stationary Distribution For each MIRU locus, let X_n be the repeat number at the n^{th} generation. There are a subset of the repeat numbers S, such that for any pair of $i, j \in S$, $p(X_{n+a} = i | X_n = j) > 0$ and $p(X_{n+b} = j | X_n = i) > 0$. This subset is called a *communication class*. We refer the largest communication class as the *major communication class* of a MIRU locus.

$\boldsymbol{\pi}^{(k)}$ of the major communication class of each MIRU loci is computed in equation (2.10). Let $\boldsymbol{q}_0^{(k)}$ be the sample distribution of the major communication class. For example, given a major communication class at MIRU locus k with states (repeat numbers) $0, 1, 2, 4$, its sample distribution is a vector with 4 entries and each entry is computed as $\dfrac{r_i}{r_0 + r_1 + r_2 + r_4}$ $(i = 0, 1, 2, 4)$ where r_i is the total appearances of repeat number i at locus k. The stationary distributions (dashed lines) together with the sample distribution, $\boldsymbol{q}_0^{(k)}$ for the major communication class (solid lines) are shown in Figures 2.9 and 2.10. Repeat numbers 'b','c' and 'd' are not in the major communication class for all the loci except MIRU 4 and they are not plotted in Figure 2.9. Also, repeat numbers 'r','s' and''t' are not in the major

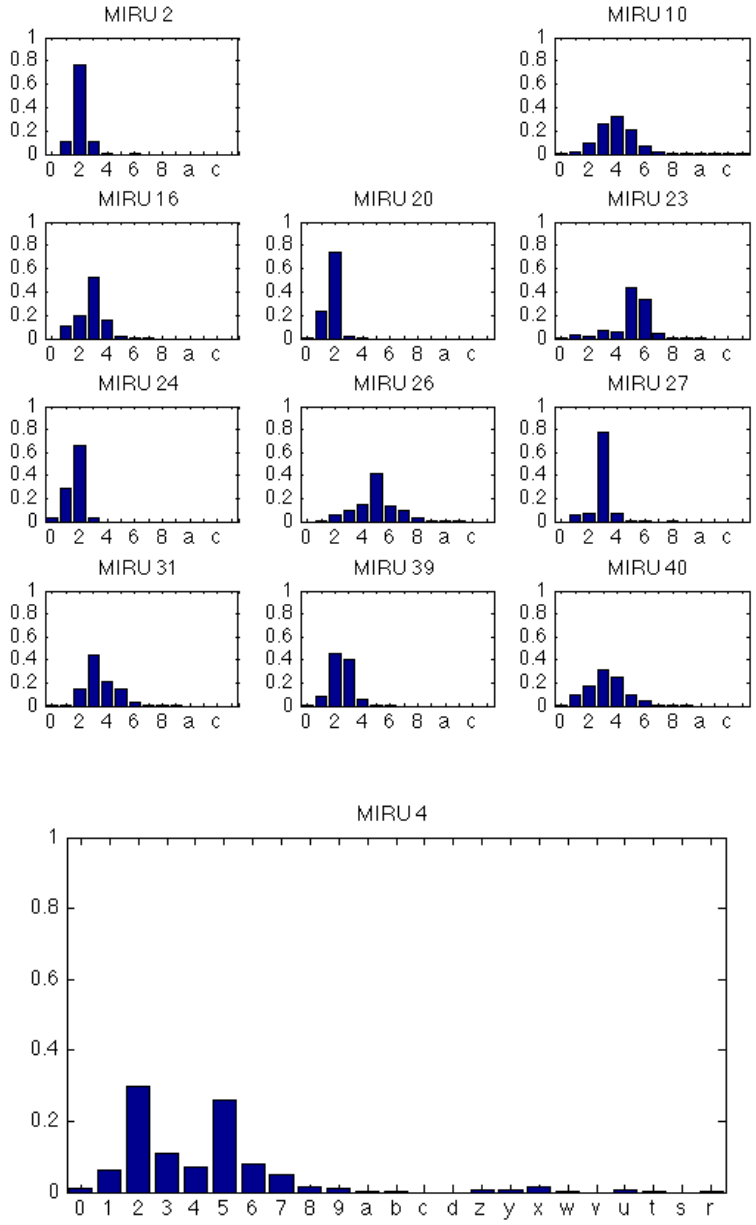


Figure 2.8: The stickyness of each repeat number measured summing up the columns of the transition count matrix while excluding the diagonal entries. The values are then normalized by the sum of these values. MIRU 2 ,20 and 24 have a sticky value at 2; MIRU 16, 27, 31, 40 at 3; MIRU 10 at 4; MIRU 26 at 5. Moreover, three loci have two sticky values: MIRU 4(2,5), MIRU 23(5,6), MIRU 39(2,3).

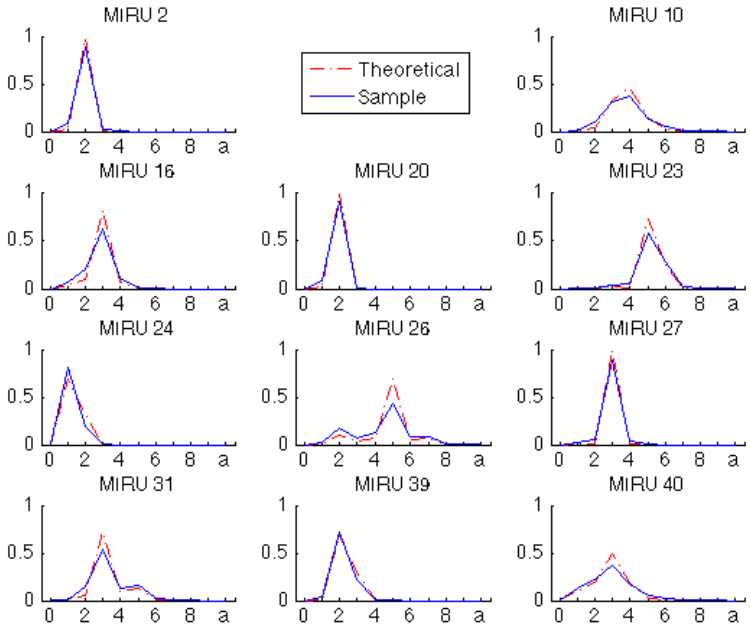


Figure 2.9: The computed theoretical stationary distribution (red dashed) and the sample distribution (solid blue) of the major class of each MIRU locus. All loci except 4 are shown in this figure.

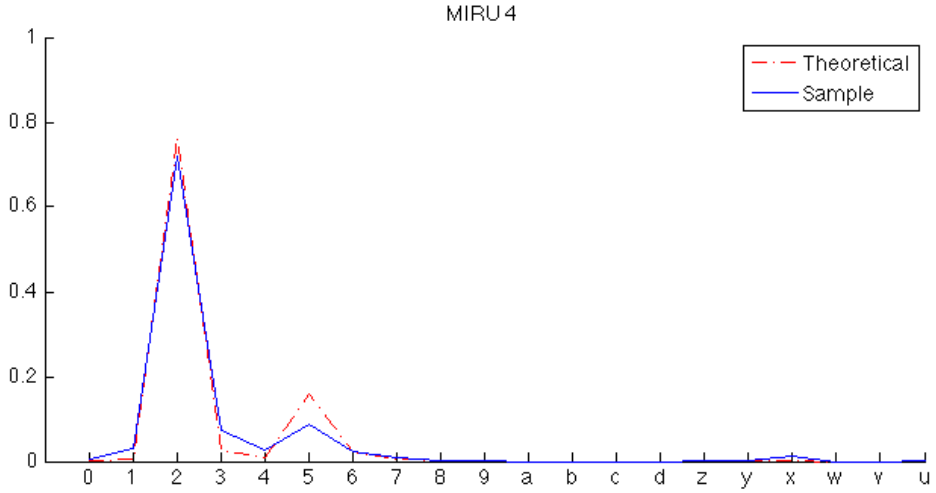


Figure 2.10: The computed theoretical stationary distribution and sample distribution of the major class of MIRU locus 4.

communication class of MIRU 4 and they thus are removed in Figure 2.10. The error of the $\boldsymbol{\pi}^{(k)}$ is studied by the Monte Carlo method. For each MIRU locus, $\tilde{P}^{(k)}$ was

simulated 10000 times with the Dirichlet distribution, as defined in equation (2.12), and the corresponding stationary distribution, $\tilde{\boldsymbol{\pi}}^{(k)}$ was computed. The standard deviation of $\tilde{\boldsymbol{\pi}}^{(k)}$'s entries are smaller than 0.016 for all loci except MIRU locus 24, whose maximum value is 0.0398, still modest compared to the its values (0.3223). The magnitude of the error in estimating $P^{(k)}$ is small, due to the large values of the transition counts. This Monte Carlo study shows that the error propagated from $\hat{P}^{(k)}$ to its stationary distribution $\boldsymbol{\pi}^{(k)}$ remain modest and our results for $\boldsymbol{\pi}^{(k)}$ are statistically significant.

The distance between the sample and stationary distribution is measured by the Kullback-Leibler divergence, which is defined as in equation (2.18). It is also known as relative entropy [38]. The values of the KL divergence of different loci are sorted in ascending order in table (2.2). Small values, such as MIRU 24: 0.0488, indicates the sample distribution is close to the stationary distribution, thus large future change in the distribution of the repeat number is not expected. Large KL divergence, such as MIRU 26: 0.1755, means the current sample distribution has not yet approached the stationary distribution, therefore, we will expect the distribution of repeat number to vary in the future.

Table 2.2: The Kullback-Leibler divergence between the sample and stationary distribution for each locus. The values are sorted from small to large

MIRU locus	24	40	39	20	10	2
KL divergence	0.0488	0.0555	0.0628	0.0690	0.0822	0.0951
MIRU locus	4	16	31	23	27	26
KL divergence	0.1014	0.1171	0.1220	0.1393	0.1449	0.1755

Rate of convergence In order to investigate convergence rate of the Markov chains, the distribution of the repeat numbers at the n^{th} generation, $\boldsymbol{q}_n^{(k)}$, were computed. The values of Kullback-Leibler (KL) divergence between the sample distributions of the repeat numbers of MIRU loci and the stationary ones for each generation, $D(\boldsymbol{q}_n^{(k)}|\boldsymbol{\pi}^{(k)})$, are plotted in Figure 2.11. Since the KL divergence of

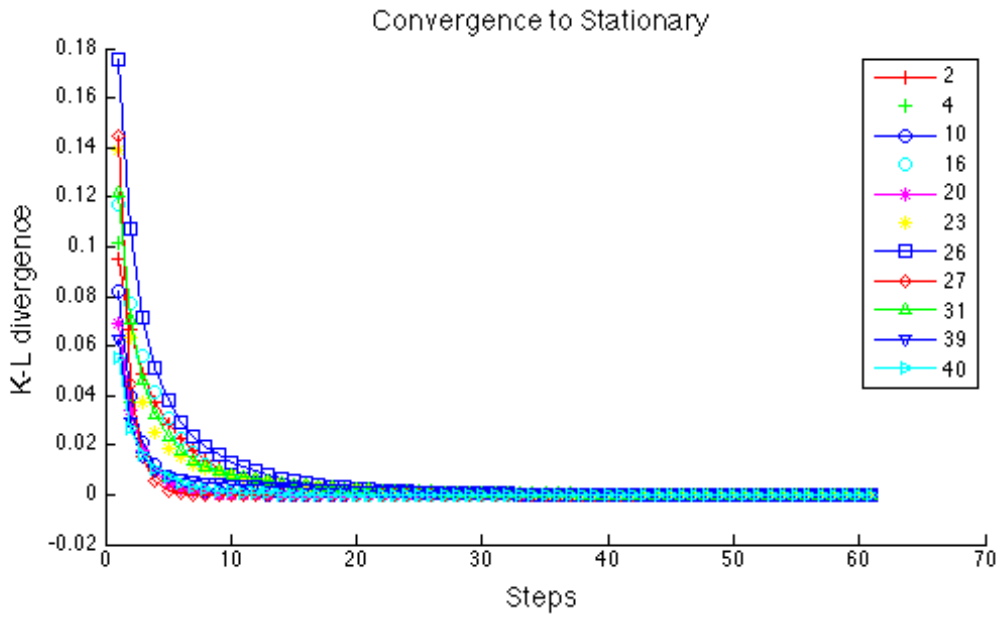


Figure 2.11: The Kullback-Leibler divergence of each MIRU loci except MIRU 24, $D(\boldsymbol{q}_{k,n}|\boldsymbol{\pi}_k)$, plotted against number of steps.

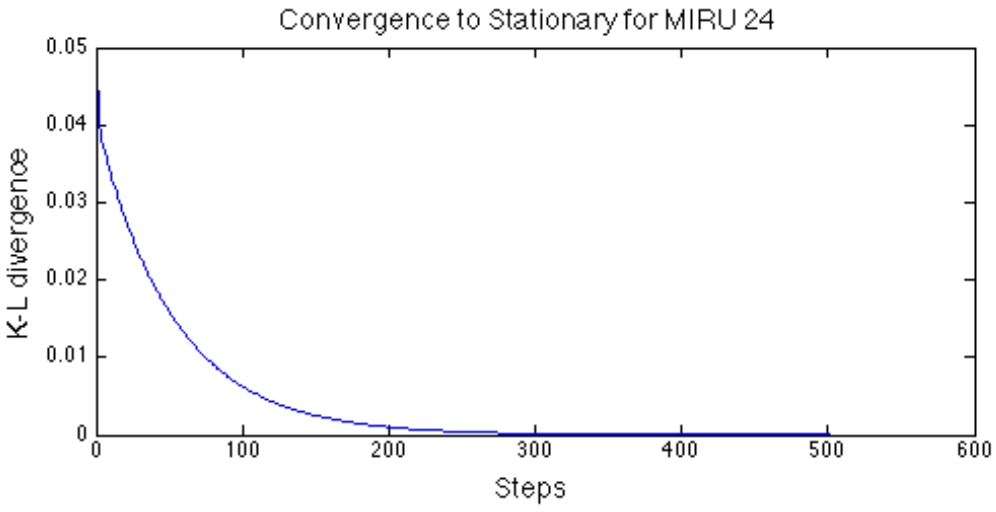


Figure 2.12: The Kullback-Leibler divergence of MIRU 24, $D(p_{24,n}|\boldsymbol{\pi}_P 24)$, plotted against number of steps.

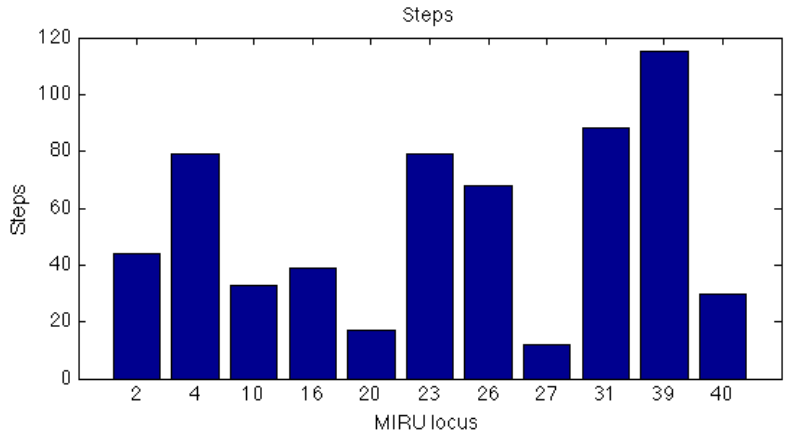


Figure 2.13: Number of steps needed for the Kullback-Leibler divergence $D(\boldsymbol{q}_{k,n}|\boldsymbol{\pi}_k)$ of each locus to approach convergence threshold $\epsilon = 1e^{-6}$, starting from sample distribution. MIRU locus 24 is excluded from this plot since it has a value of 582.

MIRU 24 decays at a rate that is significant slower than the rest of the loci, it is plotted in the separate Figure 2.12.

We chose a threshold value of $\epsilon = 1e^{-6}$ and count the number of generations for the KL divergence between $\boldsymbol{q}_n^{(k)}$ and $\boldsymbol{\pi}^{(k)}$ to drop below the threshold, i.e. $\{n|D(\boldsymbol{q}_n^{(k)}|\boldsymbol{\pi}^{(k)}) \leq \epsilon\}$. The results are plotted in Figure 2.13. MIRU 24 has the slowest convergence rate, it takes 582 steps for its KL-divergence to drop below the threshold value of $1e^{-6}$.

The different transition probability matrices for loci will give them different convergence rates. This rate is measure by $-log(|\lambda_2|)$, where λ_2 is the second largest eigenvalue of the transition probability matrix. The convergence rate for each MIRU locus is shown in Figure 2.14.

The theoretical number of steps for the 2 norm of the sample and stationary distribution of each MIRU locus is computed as in equation 2.17. The results are shown in Figure 2.15. The theoretical number of steps for the Kullback-Leibler distance of sample and stationary distribution of each MIRU locus is computed as in equation 2.20. The results are shown in Figure 2.16. Figure 2.15 and 2.16 are consistent with the results from the actual forward simulation (Figure 2.13).

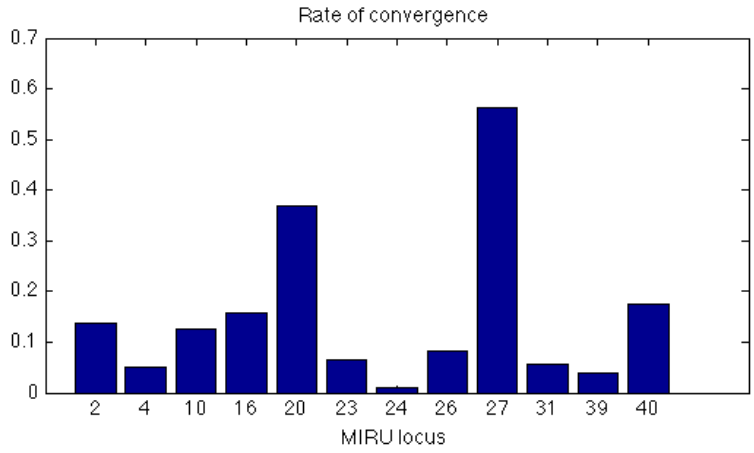


Figure 2.14: Let λ_2 be the second largest eigenvalue of the transition probability matrix for each MIRU locus. $-\ln(|\lambda_2|)$ can be used to measure the convergence rate to the stationary distribution. The values of $-\ln(|\lambda_2|)$ of each locus are plotted. As shown in the figure, locus 24 has smallest value 0.0091, while locus 27 has the greatest rate of 0.5626.

2.4 Conclusion

In this chapter, we analyzed the evolution of the MTBC from the lens of Spoligotypes and MIRU profiles. Under the rules we defined, 41,604 mutations are found among 14,453 MTBC isolates, collected by CDC and Institut Pasteur. For different MIRU loci, the transition probability of repeat numbers are computed. The heat-maps of the transition probability matrices of the MIRU loci indicate that certain repeat numbers are more popular than others in the event of a mutation. For example, when a repeat number changes in a mutation, it is more likely to change to certain values than others, i.e. 2 in locus 20 and 3 in locus 27. We also find that the small repeat numbers have tendencies to increase while the big ones tend to decrease in a mutation. We also found that in the event where repeat numbers change in a mutation, it is more likely for them to change by small values then by large values

We develop a Markov chain model for the MIRU repeat numbers based on the transition probability matrices. The stationary distributions of these Markov chains are computed. By comparing the sample distributions and stationary ones of different MIRU loci, we discovered that the sample distribution of the repeat

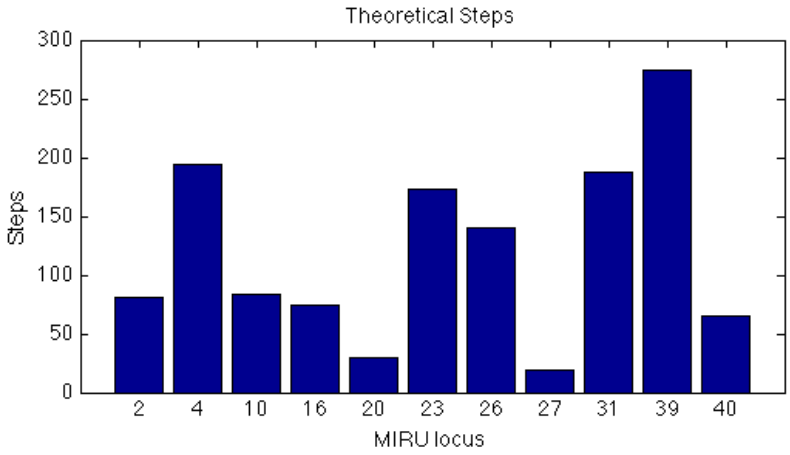


Figure 2.15: Theoretically computed number of steps needed for $\|\boldsymbol{q}_{k,n} - \boldsymbol{\pi}_k\|$ to drop within the magnitude of $\epsilon = 1e^{-6}$. The number of steps n is computed based on equation (2.17). MIRU locus 24 is excluded from this plot and it has a value of 1,293.

number of some MIRU loci are close to their stationary ones, which is measured by various methods including Kullback-Leibler (KL) divergence. For these loci, we do not expect large changes in the distributions of the repeat numbers to change in the future. On the other hand, for certain loci like 26, 27 and 4, the two distributions of their repeat numbers are different from each other base on the measurement with the KL divergence. For these loci, we expect the distribution to converge to the stationary ones.

By the analysis of the forward simulation, we investigate the convergence rate of repeat number distribution of each loci. As cited before, MIRU locus 24 corresponds to the TbD1 deletion, which is used as marker to differentiate ancestral versus modern strains. Our research shows the evidence of MTBC mutations from Modern to Ancestral strains, which is not expected. Whether this could affects the potential to use locus 24 as marker for differentiate modern and ancestral strains requires further study.

We have been investigating TB at a micro-level: studying the evolution of MTBC, the causative agent of TB. Next, we will take a step back to study the problem at a macro-level. We will spend the next few chapters discussing the probabilistic modeling of TB disease spread.

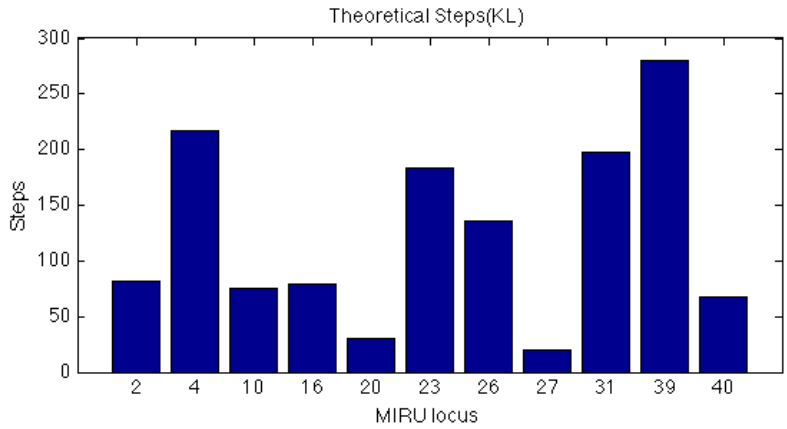


Figure 2.16: Theoretically computed number of steps needed for the Kullback-Leibler divergence $D(\boldsymbol{q}_n^{(k)}|\boldsymbol{\pi}_k)$ to drop within the magnitude of $\epsilon = 1e^{-6}$. The number of steps $\hat{n}$ is computed base on equation (2.20). MIRU locus 24 is excluded from this plot and it has a value of 1,246.

CHAPTER 3
TB Spread Modeling

3.1 Introduction

To control TB, we must understand the dynamics of its transmission. Various studies have been done on this subject [9, 39, 40]. The causative agent, MTBC, is transmitted through air. Individuals with active TB bacteria in their lungs can infect others when they sneeze, cough and speak. One of the main difference between TB and other infectious diseases is that only a small portion of individuals that are infected develop progressive disease immediately. Most people, after their initial expose to the MTBC, will mount an effective immune response which prevents the bacteria from proliferating. These individuals, although carrying TB bacteria, will not show any symptoms nor will they be infectious. However, they do have a small possibility of developing active TB through endogenous reactivation or exogenous infection [7]. In our model, each individual has one of the following three statuses:

- Susceptible: Individuals without any TB infection, but susceptible to it.
- Latent: Individuals with latent TB infection are people who are infected with MTBC, but the bacteria have not yet progressed to make the hosts have active TB. These individuals are also not infectious. For every month, they have a small possibility of developing active TB. Here we assume the main factor for active TB is endogenous reactivation. Exogenous infection occurs mainly in heavily exposed and/or immunocompromised individuals [7, 39]. The U.S. has low TB incidence and exogenous infection is rare compared to endogenous reactivation, we so ignore it in latent individuals.
- Active: Individuals are infected with MTBC and have developed active TB. These people are infectious. For every time unit, they have the possibility to be treated and thus removed from the active TB patients pool.

The relations of these three types of individuals are illustrated in Figure 3.1.

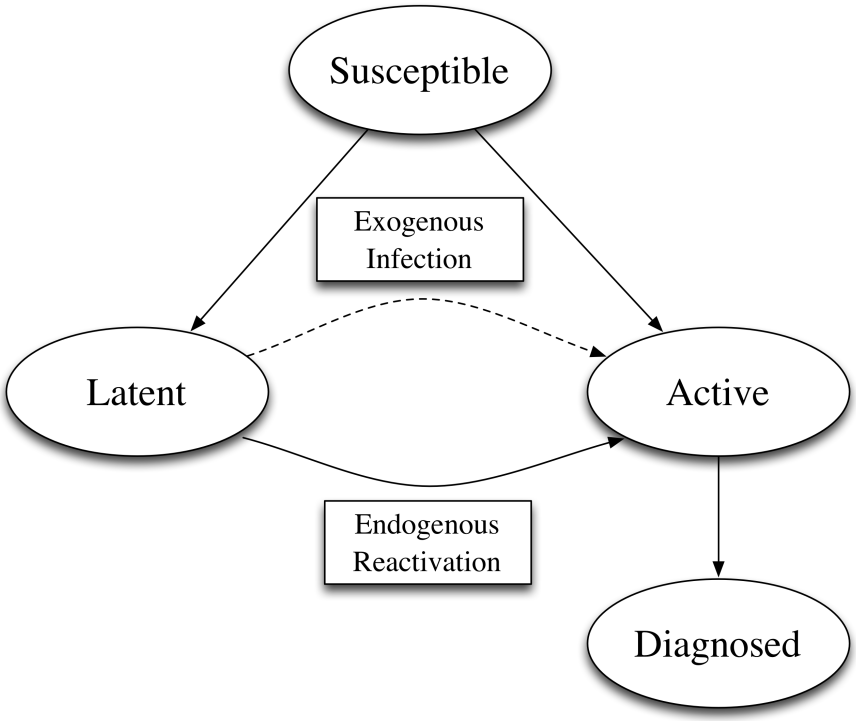


Figure 3.1: The relations of the three types of individual in the model. When a susceptible person is exogenously infected with MTBC, he/she will either become active immediately or enter the latent status. Once an individual acquires latent infection, he/she will become active through one of two ways: exogenous infection or endogenous reactivation. In this study we assume patients with latent infections progress to active TB only through endogenous reactivation.

3.2 Patient Clustering

In certain low TB incidence countries such as Australia, Canada and United Status, foreign-born persons constitute the majority of TB cases [41–43]. In the United States, 57% of reported TB cases were among foreign-born persons [11, 44], in spite of the fact that there are only 12.9% foreign-born in the total population [45]. Based on these facts, better understanding of transmission among foreign-born patients will have great impact on TB control. As mentioned before, using the DNA fingerprinting technology, MTBC isolates can be clustered into groups, within which every isolate shares the identical genotype. The MTBC isolates with identical genotypes are refered as a TB strain in this study. We define a patient cluster as

the following way.

Definition 1. *A patient cluster is defined as patients who are infected by the MTBC isolates with identical genotypes. The genotypes are defined by RFLP and spoligotypes.*

We also assume no mutations occur, i.e. no infected patient leaves the cluster by having been infected by a mutated strain.

We will focus on the groups of foreign-born TB patients within the same cluster, which normally contains 2-10 persons. In our model, we assume all the foreign-born patients develop active TB disease in one of two following ways:

1. **Endogenous reactivation**: These patients were infected before immigration but do not have active TB at the time of immigration. They develop active disease by the means of endogenous reactivation.

2. **Recent transmission**: Patients in the category entered the country without TB infection. They are exogenously infected after immigration by a TB patient in the United State and then progress to active disease.

At this stage, we exclude the possibilities that the immigrants within the cluster are infected by someone outside the cluster (domestic TB patients who share the same TB strain). We will consider this case in the next chapter.

Given the times that the foreign-born TB patients entered the country and the times when they were diagnosed, we are trying to infer whether one particular patient acquires his/her disease through endogenous reactivation or recent transmission. Being able to answer this question could help make TB control more effective, allow better allocation of scarce TB control resources, and prevent TB outbreaks. We develop mathematical models to answer this question.

3.3 Model

Mathematical epidemiological models of TB are well studied. Most of these models assume continuous time and use ordinary differential equations [7, 39, 40]. Ozcagalar et al. wrote an excellent review of the recent TB epidemiology models [27].

These models are suitable to study the dynamics among the population sizes of each type of individual (Susceptible, Latent and Active). They do a good job in finding some of the important statistics of TB, such as reproduction number and epidemic threshold [39]. These models typically do not consider MTBC genotypes. In our model, we are interested in understanding the dynamics of a smaller cluster of TB patients who are infected by the MTBC isolates with identical genotype. The small size of the patients involved allows us the develop mathematical models at an individual level. The goal of our model is to estimate the status of one individual. Ideally, we should construct a continuous time model for the accuracy. We choose to model the problem in discrete time for the following two reasons: 1) It is theoretically and technically simpler to compute with the discrete time framework; 2) The question we are trying to answer does not require the time resolution that would necessitate modeling in continuous time.

The following paragraphs will introduce the assumptions of our model. When a person is infected with TB bacteria, he/she will progress with one of the two possible routes: fast or slow route. The fast route corresponds to developing active TB immediately after the infection. The slow route represents the latent status, in which the individual carries the bacteria, but is not infectious. In the slow route, the individual will have a certain chance to develop active TB by endogenous reactivation. There is δ chance an infected person will enter the fast route, while there is $1-\delta$ chance that this person will enter the slow route. In this route, the patient will have a probability α to become actively infectious each month. Due to strict immigration medical checks before entry, new immigrants with active TB at arrival are rare, therefore we assume when a foreign-born patient enters the country, he/she is either susceptible or latent. For patients with latent MTBC infection, we assume the only way for this patient to develop active TB is through endogenous reactivation, while ignoring the possibility of exogenous infection for a person with latent status. This is because the purpose of this model is to study the epidemiology of TB in the United States, a low TB incidence country, where exogenous infections are rare relative to endogenous reactivations. [7, 39].

Suppose there are n foreign-born patients in the cluster (as in definition 1), we

label these patients $I^{(1)}$, $I^{(2)}$, ..., $I^{(n)}$. There are no people outside this cluster who can infect the patients inside the cluster. Each foreign-born patient has a entry time $t_0^{(a)}$ and a diagnosis time $t_1^{(a)}$; we use a 2 dimentional vector $\boldsymbol{t}^{(a)}$ to represent these two times, i.e. $\boldsymbol{t}^{(a)} = [t_0^{(a)}, t_1^{(a)}]$. After the diagnosis time, we assume the patient is no longer infectious and thus removed from the patient population.

The parameters used in our model are shown in the following list.

- π: The probability that one patient enters the country with latent infection.
- $\overline{\mathcal{S}^{(a)}}$: The event that the i^{th} foreign-born patient comes in with latent infection. $\mathcal{S}^{(a)}$ represents the event that the i^{th} foreign-born patient is susceptible at the time of entering.
- $\boldsymbol{t}^{(a)} = [t_0^{(a)}, t_1^{(a)}]$: The times that the i^{th} patient enters the country ($t_0^{(a)}$) and the time he/she is diagnosed ($t_1^{(a)}$).
- β: The probability that $I^{(b)}$ is infected by $I^{(a)}$ each month, given $I^{(a)}$ is infectious at that month.
- δ: The probability of entering fast route (immediately active disease) after the event of infection.
- α: The probability of developing active TB while the patient is in slow route (latent status) each month.
- γ: Once a patient develops active TB, the probability (per month) that he/she will be diagnosed each month.

Geometric random variables are used to model the time that a patient remains susceptible, latent or active. The probability mass function of the geometric random variable with the success probability θ that we use is,

$$p_X(x) = (1-\theta)^x\theta, \qquad x = 0, 1, 2, ... \tag{3.1}$$

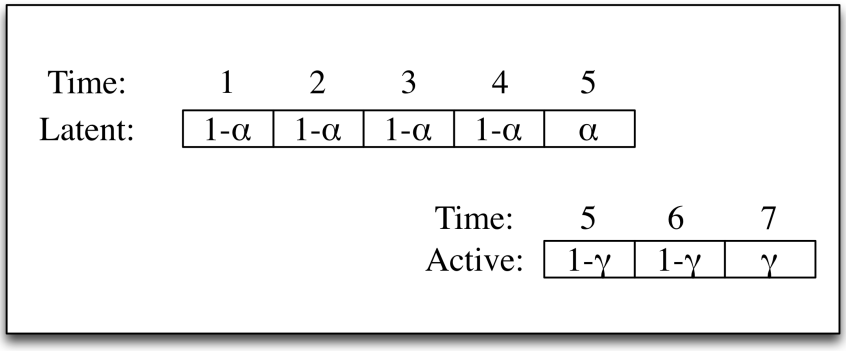


Figure 3.2: The timeline of an individual entering the country with latent infection. The patient comes in with latent infection in month 1, becomes active in month 5 and finally is diagnosed in month 7. Each month he/she remains latent with probability $1-\alpha$. Once the individual become active, each month he/she remains active with probability $1-\gamma$.

Latent Infection If an individual enters the country with latent infection, every month he/she will have a probability of α to become active. Therefore each month the individual remains latent with probability $1-\alpha$. Once the individual become active, he/she will have a probability γ to be diagnosed each month, including the month when he/she become active. Each month the individual remains active with probability $1-\gamma$. The last month the individual is diagnosed with probability γ. For example, suppose that an individual comes in with latent infection in month 1, becomes active in month 5 and is diagnosed in month 7. This happens with the following probability:

$$(1-\alpha)^4\alpha(1-\gamma)^2\gamma \tag{3.2}$$

The timeline is of this example is shown in Figure 3.2.

Susceptible If an individual is susceptible at entry time, i.e. without latent infection, and is diagnosed in a later month, this means that he/she must be infected after the entry. As discussed before, after the infection, the individual has δ chance to becomes active immediately (fast route) and probability of $1-\delta$ to become latent (slow route). For instance, let's consider a TB cluster with size 2. Suppose an in-

dividual comes in susceptible in the month 1 and is diagnosed in month 7. Just for the purpose of demonstration, assume the other patient is active from month 1 to 7. This implies that each month the susceptible individual will have a probability β to be infected. Suppose this individual is infected in the month 3 and becomes active immediately. We are computing the probability of the event in which the following happen: 1) The patient (susceptible) comes in in the 1^{st} month, AND 2) is infected in the 3^{rd} month, AND 3) becomes active in the 3^{rd} month, AND 4) is diagnosed in the 7^{th} month. This probability is written as follows,

$$\delta(1-\beta)^2\beta(1-\gamma)^4\gamma \tag{3.3}$$

Note that the month when the individual is infected is counted as one month that he/she remains active.

Instead of being active immediately after infection, suppose the patient becomes latent and later active in month 5. While cases 1), 2) and 4) in the previous event remain the same, case 3) in the previous paragraph is now changed to "becomes active in the 5^{th} month". The corresponding probability is computed as:

$$(1-\delta)(1-\beta)^2\beta(1-\alpha)^2\alpha(1-\gamma)^2\gamma \tag{3.4}$$

Note that the month when the patient is infected is counted as one month he/she remains latent. Similarly, the month the individual become active is counted as one month he/she remains active. The timeline of the susceptible case is shown in Figure 3.3.

2-person case For simplicity, we will start introducing our model with a cluster of two patients: $I^{(1)}$ and $I^{(2)}$. We assume there are no active TB patients outside the cluster who could infect $I^{(1)}$ and $I^{(2)}$. This implies that there are at least one of $I^{(1)}$ and $I^{(2)}$ enters the country with latent infection. Therefore, we only consider the following three cases: $\overline{\mathcal{S}^{(1)}}\,\overline{\mathcal{S}^{(2)}}$, $\overline{\mathcal{S}^{(1)}}\mathcal{S}^{(2)}$ and $\mathcal{S}^{(1)}\overline{\mathcal{S}^{(2)}}$. We would like to compute the conditional probability that $I^{(2)}$ being latently infected at the time of entry given the entry and diagnosis times of $I^{(1)}$ and $I^{(2)}$.

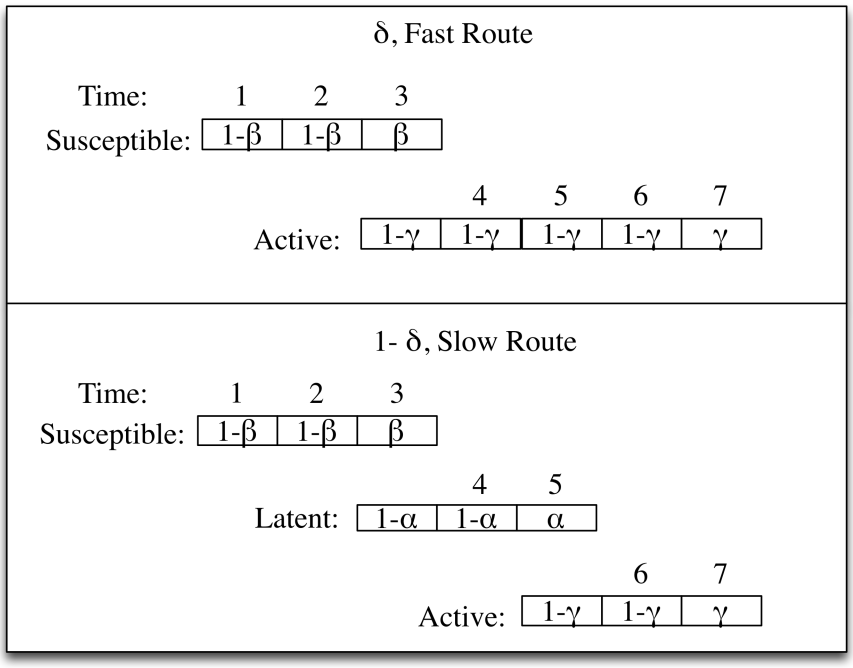


Figure 3.3: The timeline of an individual who is susceptible at entry time. He/she comes in month 1, is infected in month 3 and is diagnosed in month 7. Assume the other patient in the 2-person cluster is active from month 1 to 7. With probability δ, the individual will enter the fast route. With probability $1-\delta$, the person will enter the slow route.

We will present two methods: *2-body method* which computes the probability of every possible case; *1-body mean field method* which makes simplifying assumptions.

3.3.1 2-body Method

The 2-body method is the brute force computation of all the cases. The conditional probabilities of observing the diagnosis times of I_1 and I_2 given their entry times and the initial statuses are computed. The probability that I_2 enters with latent infection can be computed using these conditional probabilities.

i. $\overline{\mathcal{S}^{(1)}}\ \overline{\mathcal{S}^{(2)}}$ Both patients had latent infection at the time of immigration. They progress to active TB through slow route. The time that a patient remain latent can be modeled by a geometric random variable with success probability α. Once active, the time that he/she will remain active until getting diagnosed can be modeled by

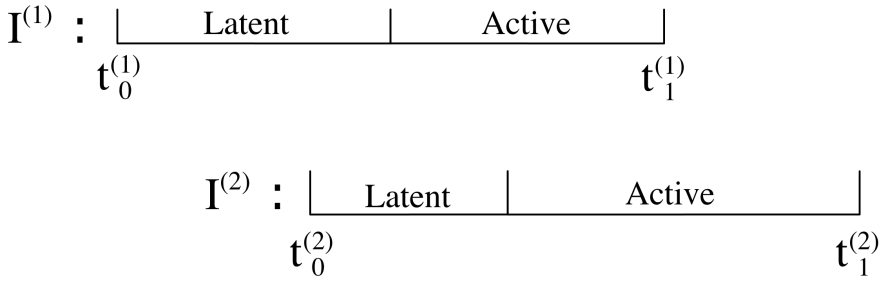


Figure 3.4: The time each patient remain latent will be a geometric random variable with success probability α, the time he/she remains active will be another geometric random variable with success probability γ. For each patient, the time he/she spends between entering to getting diagnosed will be a sum of two geometric random variables

.

another geometric random variable with success probability γ. Thus for each patient, the length of time from entering the country until getting diagnosed can be viewed as a sum of two geometric random variables as illustrated in figure 3.4.

For our derivation, we require the following simple result. Give two geometric random variables X_1 and X_2 with success probability θ_1 and θ_2, the probability of observing the event of $X_1 + X_2 = x$ will be

$$\begin{aligned} p(X_1 + X_2 = x) &= \sum_{i=0}^{x} (1-\theta_1)^i \theta_1 (1-\theta_2)^{x-i} \theta_2 \\ &= \frac{\theta_1\theta_2[(1-\theta_1)^{x+1} - (1-\theta_2)^{x+1}]}{\theta_2 - \theta_1} \end{aligned} \tag{3.5}$$

Given a patient has latent infection and he/she entered at $t_0^{(a)}$ and was diagnosed at $t_1^{(a)}$, the total length of time each individual spends in the country is $t_1^{(a)} - t_0^{(a)}$, excluding the last month when he/she is diagnosed. This value is modeled by a sum of two geometric random variables with success probability α and γ respectively. Therefore, the corresponding probability follows equation (3.5),

$$p(t_1^{(a)} | t_0^{(a)}, \overline{\mathcal{S}^{(a)}}) = \frac{\alpha\gamma[(1-\alpha)^{t_1^{(a)} - t_0^{(a)} + 1} - ((1-\gamma)^{t_1^{(a)} - t_0^{(a)} + 1})]}{\gamma - \alpha} \tag{3.6}$$

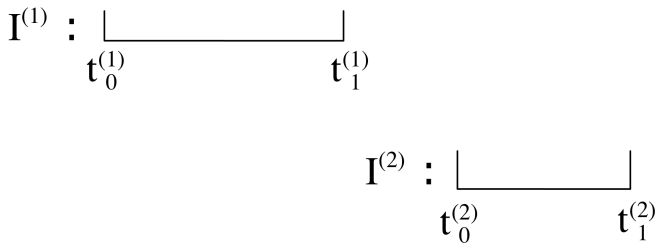


Figure 3.5: An illustration of the time periods that $I^{(1)}$ and $I^{(2)}$ spend from entry to diagnosis. Note that these two periods do not overlap.

Note that in the case, one patient's entry time is independent of the other patient's diagnosis time and entry status. Therefore, we have the following,

$$\begin{aligned} & p(t_1^{(1)}|t_0^{(1)}, \overline{\mathcal{S}^{(1)}})p(t_1^{(2)}|t_0^{(2)}, \overline{\mathcal{S}^{(2)}}) \\ = \; & p(t_1^{(1)}|t_0^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(1)}}, \overline{\mathcal{S}^{(2)}})p(t_1^{(2)}|t_0^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(1)}}, \overline{\mathcal{S}^{(2)}}) \\ = \; & p(t_1^{(1)}, t_1^{(2)}|t_0^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(1)}}, \overline{\mathcal{S}^{(2)}}) \end{aligned}$$

ii. $\overline{\mathcal{S}^{(1)}}\mathcal{S}^{(2)}$ $I^{(1)}$ has latent infection and $I^{(2)}$ is susceptible at the time of immigration. If the time periods that $I^{(1)}$ and $I^{(2)}$ each spend between entry and diagnosis do not overlap, which means $I^{(1)}$ enters after $I^{(2)}$ is diagnosed ($t_0^{(1)} > t_1^{(2)}$) or $I^{(2)}$ enters after $I^{(1)}$ is diagnosed ($t_1^{(1)} < t_0^{(2)}$), then we have $p(\boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)}, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)}) = 0$. Figure 3.5 illustrates this case.

If the two time periods overlap, after $I^{(1)}$ becomes active, $I^{(2)}$ will start to have the risk to be infected. If $I^{(1)}$ successfully infected $I^{(2)}$ by the time he/she is diagnosed, $I^{(2)}$ will enter one of the two routes: fast route with probability δ or slow route with $1-\delta$. In the fast route, $I^{(2)}$ becomes active immediately and remains so until getting diagnosed. On the other hand, if $I^{(2)}$ enters the slow route, he/she will have latent infection. $I^{(2)}$ will remain latent until becoming active. Finally, $I^{(2)}$ will be diagnosed and removed from the population. This case is illustrated by Figure 3.6.

Similar to the previous case, the time that a patient remains latent or active will be modeled by geometric random variables with success probabilities α and γ

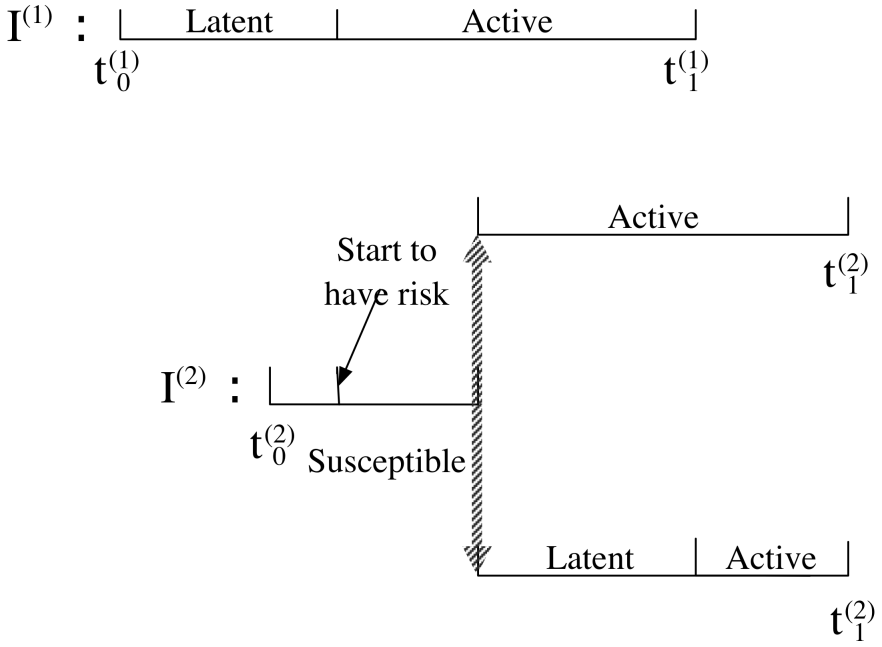


Figure 3.6: Similar to the $\overline{\mathcal{S}^{(1)}}$,$\overline{\mathcal{S}^{(2)}}$ case, the times that each patient remains latent (active) will be modeled by geometric random variables with success probabilities α (γ). The time from when $I^{(2)}$ starts having the risk to be infected to the time of infection, will be modeled by geometric random variable with success probability β. After infection, $I^{(2)}$ will enter one of the two routes.

respectively. The time when $I^{(2)}$ starts having the risk to be infected will be $I^{(1)}$'s active time or $I^{(2)}$'s entry time, whichever is later. The time when $I^{(2)}$ stops having the risk will be $I^{(1)}$ or $I^{(2)}$'s diagnosis time, whichever is earlier. Suppose the time when $I^{(1)}$ becomes active is t_{active}, the period that $I^{(2)}$ is at risk to the infected is

$$\max(t_{active}, t_0^{(2)}) \rightarrow \min(t_1^{(1)}, t_1^{(2)}) \tag{3.7}$$

Given that $I^{(1)}$ is active at a certain month, assuming $I^{(2)}$ already entered the country, the probability that $I^{(2)}$ is infected will be β. Therefore, the length of the time from when $I^{(2)}$ starts to have risk to the time of infection will be modeled by a geometric random variable with success probability β. This case is illustrated in Figure 3.6.

Again, the entering and diagnosing times of $I^{(1)}$ and $I^{(2)}$ are represented by vectors $\boldsymbol{t}^{(1)} = [t_0^{(1)}, t_1^{(1)}]$, $\boldsymbol{t}^{(2)} = [t_0^{(2)}, t_1^{(2)}]$. Given $I^{(1)}$ has latent infection and $I^{(2)}$ is susceptible at the time of entering and their entry times are $t_0^{(1)}$, $t_0^{(2)}$, the probability of observing the diagnosis times $t_1^{(1)}$, $t_1^{(2)}$ will be

$$\begin{aligned} p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)}) &= \sum_{i=t_0^{(1)}}^{t_1^{(1)}} \Bigg\{ (1-\alpha)^{i-t_0^{(1)}} \alpha (1-\gamma)^{t_1^{(1)}-i} \gamma \\ &\cdot \sum_{j=\tau_1}^{\tau_2} (1-\beta)^{j-\tau_1} \beta \cdot \Big\{ \delta (1-\gamma)^{t_1^{(2)}-j} \gamma \\ &+ (1-\delta) \frac{\alpha\gamma[(1-\alpha)^{t_1^{(2)}-j+1} - ((1-\gamma)^{t_1^{(2)}-j+1})]}{\gamma - \alpha} \Big\} \Bigg\} \end{aligned} \tag{3.8}$$

where $\tau_1 = \max(i, t_0^{(2)})$ and $\tau_2 = \min(t_1^{(1)}, t_1^{(2)})$.

Explanation of Equation (3.8)

- $\sum_{i=t_0^{(1)}}^{t_1^{(1)}} (1-\alpha)^{i-t_0^{(1)}} \alpha (1-\gamma)^{t_1^{(1)}-i} \gamma$. This part sums over the probabilities of all the months, i, when $I^{(1)}$ could become active. Each month while $I^{(1)}$ has latent infection, there is α chance of becoming active. Once activated, $I^{(1)}$ will have a γ chance of being diagnosed each month. The month that $I^{(1)}$ becomes active is counted as a month that he/she is active and can be diagnosed; therefore we have $t_1^{(1)} - i$ as superscript.
- $\sum_{j=\tau_1}^{\tau_2} (1-\beta)^{j-i} \beta$. The second part sums over the probabilities of all the months, when $I^{(2)}$ will be infected. We are assuming in the month when $I^{(1)}$ is diagnosed, which is $t_1^{(1)}$ here, he/she can still infect $I^{(2)}$. Given the activation time of $I^{(1)}$ is i, from the outer summation, the months that $I^{(2)}$ could be infected range from $\tau_1 = \max(i, t_0^{(2)})$ to $\tau_2 = \min(t_1^{(1)}, t_1^{(2)})$. Each of these months, the probability of $I^{(2)}$ being infected is β.

- $\delta(1-\gamma)^{t_1^{(2)}-j}\gamma$. Once $I^{(2)}$ is infected at month j, as discussed before, there is a δ chance that he/she will enter the fast route, i.e. becoming active immediately. In this route, the patient will have a γ chance being diagnosed each month.
- $(1-\delta)\dfrac{\alpha\gamma[(1-\alpha)^{t_1^{(2)}-j+1}-((1-\gamma)^{t_1^{(2)}-j+1})]}{\gamma-\alpha}$. $I^{(2)}$ could also enter the slow route, i.e. the patient progresses from latent to active and is finally diagnosed at $t_1^{(2)}$. The length of time from the moment $I^{(2)}$ enters the slow route (month j) to the moment before he/she is diagnosed (month $t_1^{(2)}$) is $t_1^{(2)}-j$. This is modeled as a sum of two geometric random variables with success probabilities α and γ. We can use equation (3.5) again to compute this probability.

Under the current assumption of constant β, the inner summation of equation (3.8) are combinations of the geometric sums. After simplification, it becomes the following

$$\begin{aligned} p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)}) &= \sum_{i=t_0^{(1)}}^{t_1^{(1)}} (1-\alpha)^{i-t_0^{(1)}} \alpha (1-\gamma)^{t_1^{(1)}-i}\gamma \\ &\cdot \left\{ \frac{\delta\beta\gamma}{\beta-\gamma}A + \frac{(1-\delta)\alpha\beta\gamma}{\gamma-\alpha}\left[\frac{1}{\beta-\alpha}B - \frac{1}{\beta-\gamma}C\right]\right\} \end{aligned} \tag{3.9}$$

where A, B and C are defined as follows,

$$\begin{aligned} A &= (1-\gamma)^{t_1^{(2)}-\tau_1+1} - (1-\beta)^{\tau_2-\tau_1+1}(1-\gamma)^{t_1^{(2)}-\tau_2} \\ B &= (1-\alpha)^{t_1^{(2)}-\tau_1+2} - (1-\beta)^{\tau_2-\tau_1+1}\beta(1-\alpha)^{t_1^{(2)}-\tau_2+1} \\ C &= (1-\gamma)^{t_1^{(2)}-\tau_1+2} - (1-\beta)^{\tau_2-\tau_1+1}\beta(1-\gamma)^{t_1^{(2)}-\tau_2+1} \end{aligned}$$

Equation (3.9) can be further simplified using the formula of geometric sum. However, we will have to use Equation (3.8) when we considering the more important case later, where β is time-dependent. Further simplifying Equation (3.9) will create too complicated an expression to be of much practical use.

iii. $\mathcal{S}^{(1)}\overline{\mathcal{S}^{(2)}}$ This case will be exactly the same with $\overline{\mathcal{S}^{(1)}}\mathcal{S}^{(2)}$, after switching the indexes.

iv. $\mathcal{S}^{(1)}\mathcal{S}^{(2)}$ At this stage, we assume there is no active patients other than $I^{(1)}$ and $I^{(2)}$. Therefore, we have

$$p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \mathcal{S}^{(1)}, \mathcal{S}^{(2)}) = 0 \tag{3.10}$$

Now we have the conditional probabilities of the diagnosis time given the entry time and the initial statuses and we would like to compute $p(\overline{\mathcal{S}^{(2)}} | t_0^{(1)}, t_0^{(2)}, t_1^{(1)}, t_1^{(2)})$. Since $I^{(1)}$ and $I^{(2)}$ are the first and second patients who are diagnosed in the foreign born population, we also need to include the information of "no other people are diagnosed by $I^{(2)}$'s diagnosis time. Let G_α, G_β, G_γ be the geometric random variables with success probability α, β, γ respectively. Note that here β is constant and represents the average probability of a susceptible person being infected in a particular month. Let X_L, X_S be the time latent, susceptible people spend from entry to diagnosis. Based on these assumptions, we have $X_L = G_\alpha + G_\gamma$; $X_S = G_\beta + G_\gamma$ with probability δ, and $X_S = G_\beta + G_\alpha + G_\gamma$ with probability $1-\delta$. Since $\gamma >> \alpha, \beta$, we have $G_\alpha, G_\beta >> G_\gamma$. We therefore approximate $X_L = G_\alpha$; $X_S = G_\beta$ with probability δ, and $X_S = G_\beta + G_\alpha$ with probability $1 - \delta$. Let $F_{X_L}(x)$ be the cumulative distribution function of X_L, i.e. $F_{X_L}(x) = P\{X_L \leq x\}$. $1 - F_{X_L}(x)$ is easy to obtain,

$$1 - F_{X_L}(x) = (1-\alpha)^{x+1} \tag{3.11}$$

To derive $F_{X_S}(x)$, we need to formulate $P\{G_\beta + G_\alpha > x\}$ first. The probability mass function of the sum of two geometric random variables is defined as in Equation (3.5):

$$P\{G_\beta + G_\alpha = x\} = \frac{\beta\alpha\left[(1-\beta)^{x+1} - (1-\alpha)^{x+1}\right]}{\alpha - \beta} \tag{3.12}$$

$$\begin{aligned} P\{G_\beta + G_\alpha > x\} &= P\{G_\beta + G_\alpha = x+1\} + P\{G_\beta + G_\alpha = x+2\} + ... \\ &= \frac{\alpha(1-\beta)^{x+2} - \beta(1-\alpha)^{x+2}}{\alpha - \beta} \end{aligned} \tag{3.13}$$

Again, let $F_{X_S}(x)$ be the cumulative distribution function of X_S. Based on the result from Equation (3.13), $1 - F_{X_S}(x)$ can be computed as follows:

$$\begin{aligned} 1 - F_{X_S}(x) &= \delta(1-\beta)^{x+1} + (1-\delta)\frac{\alpha(1-\beta)^{x+2} - \beta(1-\alpha)^{x+2}}{\alpha - \beta} \\ &= \left[\delta + \frac{(1-\delta)(1-\beta)\alpha}{\alpha - \beta}\right](1-\beta)^{x+1} - \frac{(1-\delta)\beta}{\alpha - \beta}(1-\alpha)^{x+2} \end{aligned} \tag{3.14}$$

Let T be the time that an individual spends from entry to $t_1^{(2)}$, i.e. the individual has not yet been diagnosed by $t_1^{(2)}$. Assume the average life expectancy of the immigrants (starting from the time of first entry to the country) is 600 months. In the simulation, every immigrant is removed from the population after 600 months after entry. Assume people enter with constant rate, after sufficiently long time, T is uniformly distributed in $[0, 599]$. The expectations of $1 - F_{X_L}(T)$ and $1 - F_{X_S}(T)$ are computed as follows

$$\mathbb{E}[1 - F_{X_L}(T)] = \frac{(1-\alpha)^1 - (1-\alpha)^{601}}{600\alpha} \tag{3.15}$$

$$\mathbb{E}[1 - F_{X_S}(T)] = A\frac{(1-\beta)^1 - (1-\beta)^{601}}{600\beta} - B\frac{(1-\alpha)^2 - (1-\alpha)^{602}}{600\alpha} \tag{3.16}$$

Where

$$A = \delta + \frac{(1-\delta)(1-\beta)\alpha}{\alpha - \beta}$$
$$B = \frac{(1-\delta)\beta}{\alpha - \beta}$$

For convenience, let's use the following notations

- $\mathbb{E}[1 - F_{X_L}(T)] \to H_{X_L}$
- $\mathbb{E}[1 - F_{X_S}(T)] \to H_{X_S}$

H_{X_L} (H_{X_S}) computes the average probability of an person with entry status as

latent (susceptible) and remains not being diagnosed for a random time T, which is the time from his/her entry to $I^{(2)}$'s diagnosis time. Assume the initial disease status is independent of the entry times, we have the following

- $P_{LL} = p(t_1^{(1)}, t_1^{(2)}, \overline{\mathcal{S}^{(1)}}, \overline{\mathcal{S}^{(2)}} | t_0^{(1)}, t_0^{(2)}) = p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(1)}}, \overline{\mathcal{S}^{(2)}}) \pi^2$
- $P_{SS} = p(t_1^{(1)}, t_1^{(2)}, \mathcal{S}^{(1)}, \mathcal{S}^{(2)} | t_0^{(1)}, t_0^{(2)}) = p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \mathcal{S}^{(1)}, \mathcal{S}^{(2)}) (1-\pi)^2$
- $P_{SL} = p(t_1^{(1)}, t_1^{(2)}, \mathcal{S}^{(1)}, \overline{\mathcal{S}^{(2)}} | t_0^{(1)}, t_0^{(2)}) = p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \mathcal{S}^{(1)}, \overline{\mathcal{S}^{(2)}}) \pi(1-\pi)$
- $P_{LS} = p(t_1^{(1)}, t_1^{(2)}, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)} | t_0^{(1)}, t_0^{(2)}) = p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)}) \pi(1-\pi)$

Assume there are N_L latent people and N_S susceptible persons in our problem universe. In the first scenario $I^{(1)}$ and $I^{(2)}$ are both latent at entry, the average probability that the rest of the people with latent infection have not been diagnosed by $t_1^{(2)}$ is $H_{X_L}^{N_L-2}$. The probability that all the N_S susceptible people have not been infected and diagnosed by T is $H_{X_S}^{N_S}$. Therefore $P_{LL} H_{X_L}^{N_L-2} H_{X_S}^{N_S}$ computes the probability that the first two patients are diagnosed at $t_1^{(1)}$ and $t_1^{(2)}$, had latent infections at the times of entry and the rest of people have not been diagnosed by $t_1^{(2)}$. Similar to $P_{LL} H_{X_L}^{N_L-2} H_{X_S}^{N_S}$, $P_{SL} H_{X_L}^{N_L-1} H_{X_S}^{N_S-1}$ computes the case where $I^{(1)}$ was susceptible and $I^{(2)}$ was latent; $P_{LS} H_{X_L}^{N_L-1} H_{X_S}^{N_S-1}$ computes the case where $I^{(1)}$ was latent and $I^{(2)}$ was susceptible; $P_{SS} H_{X_L}^{N_L} H_{X_S}^{N_S-2}$ computes the case where both patients were susceptible. Let us denote these expressions as follows,

- $P_{LL} H_{X_L}^{N_L-2} H_{X_S}^{N_S}$ as $\tilde{p}(t_1^{(1)}, t_1^{(2)}, \overline{\mathcal{S}^{(1)}}, \overline{\mathcal{S}^{(2)}} | t_0^{(1)}, t_0^{(2)})$
- $P_{SL} H_{X_L}^{N_L-1} H_{X_S}^{N_S-1}$ as $\tilde{p}(t_1^{(1)}, t_1^{(2)}, \mathcal{S}^{(1)}, \mathcal{S}^{(2)} | t_0^{(1)}, t_0^{(2)})$
- $P_{LS} H_{X_L}^{N_L-1} H_{X_S}^{N_S-1}$ as $\tilde{p}(t_1^{(1)}, t_1^{(2)}, \mathcal{S}^{(1)}, \overline{\mathcal{S}^{(2)}} | t_0^{(1)}, t_0^{(2)})$
- $P_{SS} H_{X_L}^{N_L} H_{X_S}^{N_S-2}$ as $\tilde{p}(t_1^{(1)}, t_1^{(2)}, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)} | t_0^{(1)}, t_0^{(2)})$

The conditional probability of $I^{(2)}$ having latent infection at entry given all the

timing information of both patients is computed as follows,

$$\tilde{p}(\overline{\mathcal{S}^{(2)}}|t_0^{(1)},t_0^{(2)},t_1^{(1)},t_1^{(2)}) = \frac{\tilde{p}(t_1^{(1)},t_1^{(2)},\overline{\mathcal{S}^{(2)}}|t_0^{(1)},t_0^{(2)})}{\tilde{p}(t_1^{(1)},t_1^{(2)}|t_0^{(1)},t_0^{(2)})} = \frac{P_{LL}H_{X_L}^{N_L-2}H_{X_S}^{N_S} + P_{SL}H_{X_L}^{N_L-1}H_{X_S}^{N_S-1}}{P_{LL}H_{X_L}^{N_L-2}H_{X_S}^{N_S} + P_{SL}H_{X_L}^{N_L-1}H_{X_S}^{N_S-1} + P_{LS}H_{X_L}^{N_L-1}H_{X_S}^{N_S-1} + P_{SS}H_{X_L}^{N_L}H_{X_S}^{N_S-2}} \tag{3.17}$$

Divide both the numerator and denominator by $H_{X_L}^{N_L-2}H_{X_S}^{N_S-2}$ and we have the following:

$$\begin{aligned} & \tilde{p}(\overline{\mathcal{S}^{(2)}}|t_0^{(1)},t_0^{(2)},t_1^{(1)},t_1^{(2)}) \\ = & \frac{P_{LL}H_{X_S}^2 + P_{SL}H_{X_L}H_{X_S}}{P_{LL}H_{X_S}^2 + P_{SL}H_{X_L}H_{X_S} + P_{LS}H_{X_L}H_{X_S} + P_{SS}H_{X_L}^2} \end{aligned} \tag{3.18}$$

With results from equation (3.18), the conditional probability of $I^{(2)}$ being susceptible at entry can be easily computed.

$$\begin{aligned} & \tilde{p}(\mathcal{S}^{(2)}|t_0^{(1)},t_0^{(2)},t_1^{(1)},t_1^{(2)}) \\ = & 1 - \tilde{p}(\overline{\mathcal{S}^{(2)}}|t_0^{(1)},t_0^{(2)},t_1^{(1)},t_1^{(2)}) \\ = & \frac{P_{SS}H_{X_L}^2 + P_{LS}H_{X_L}H_{X_S}}{P_{LL}H_{X_S}^2 + P_{LS}H_{X_L}H_{X_S} + P_{LS}H_{X_L}H_{X_S} + P_{SS}H_{X_L}^2} \end{aligned} \tag{3.19}$$

3.3.2 1-body Mean Field Method

The previous model works well with two patients. When the number of patients increases, however, the model becomes complicated quickly since it needs to compute all the possible infection scenarios. For instance, assume there are three patients, $I^{(1)}$, $I^{(2)}$ and $I^{(3)}$, where $I^{(1)}$ and $I^{(2)}$ entered susceptible and $I^{(3)}$ entered with latent infection, which is $\mathcal{S}^{(1)}\mathcal{S}^{(2)}\overline{\mathcal{S}^{(3)}}$ in our notation. We need to consider whether both of $I^{(1)}$ and $I^{(2)}$ were infected by $I^{(3)}$ or if $I^{(2)}$ was infected by $I^{(3)}$ first and then $I^{(1)}$ was infected by $I^{(2)}$, etc.

To solve this problem, we present a 1-body mean field method. Since we are

only interested in how one of the patients acquired TB infection, we are computing for more information than we need in the **1-body** method. In reality, there are n patients $I^{(1)}$ to $I^{(n)}$, but we are only interested in knowing one patient's status. Let us say this patient is $I^{(b)}$. For the purpose of inferring $I^{(b)}$'s disease status at immigration, we approximate the problem by ignoring the cases in which $I^{(b)}$ infects others.

First, we define $\Gamma_a(k)$ as the infectivity contributed by $I^{(a)}, a = 1...n, a \neq b$ in month k. Let N_F be the average total foreign born population in our modeling universes, the probability that $I^{(b)}$ is infected by $I^{(a)}$ in one particular month k is defined to be $\beta_a(k) = \dfrac{\Gamma_a(k)}{N_F}$. Note that $\Gamma_a(k)$ is the infectivity contributed by one active TB patient. An susceptible patient in the cluster will experience this infectivity so that have a probability $\beta_a(k)$ to be infected by the active patient. If $I^{(b)}$ enters with latent infection, the model does not need the information of the other $n-1$ patients. In the case where $I^{(b)}$ enters susceptible, we need to know the probability of being infected by the other $n-1$ patients in a particular month k. Instead of computing exactly which patients are active in month k and taking average over all the possible combinations, for each patient among $\{I^{(a)}\}_{a=1...n,a\neq b}$, we compute the probability that he/she is active in month k given the diagnosis time. The probability of being infected by $I^{(a)}$ in month k, $\beta_a(k)$, will be $\beta \cdot p\{I^{(a)}$ is active in month $k|I^{(a)}$is diagnosed at $t_1^{(a)}\}$. The probability being infected by anyone of the $n-1$ patients in month k will be computed as $1-\prod_{a=1,a\neq b}^{n}[1-\beta_a(k)]$

Let $A^{(a)}$ be the number of months that the a^{th} patient spends from the month he/she becomes active to the month before getting diagnosed. The likelihood that $I^{(a)}$ is infectious at a certain month k is assumed to geometrically decay according to the length of time from k to $t_1^{(a)}$. We assume $A^{(a)}$ follows a geometric distribution with success probability γ. Note that γ is also the probability that a patient is diagnosed per month while he/she is in the active TB status. Here we reverse the process to capture the dynamics of how long a patient remains infectious prior to his/her diagnosis time.

The probability $I^{(a)}$ is infectious in month k (given $t_0^{(a)} \leq k \leq t_1^{(a)}$) is equivalent to the probability of $A^{(a)} \geq t_i^{(1)} - k$, which is computed as $(1-\gamma)^{t_1^{(a)}-k}$. This is because

$A^{(a)} \geq t_1^{(a)} - k$ means there are at least $t_1^{(a)} - k$ months at which the active $I^{(a)}$ is not diagnosed. For convenience, let us denote:

$$p\{I^{(a)} \text{ is active in month } k | I^{(a)} \text{is diagnosed at } t_1^{(a)}\} \text{ as } p_a(t_0^{(a)}, t_1^{(a)}, k)$$

and it is computed as follows:

$$p_a(t_0^{(a)}, t_1^{(a)}, k) = \begin{cases} (1-\gamma)^{t_1^{(a)}-k} & \text{if } t_0^{(a)} \leq k \leq t_1^{(a)} \\ 0 & \text{otherwise} \end{cases} \tag{3.20}$$

For simplicity, let's start from the 2 patient case. Let two foreign-born patients $I^{(1)}$, $I^{(2)}$, enter the country at $t_0^{(1)}$, and $t_0^{(2)}$. They are diagnosed at $t_1^{(1)}$, $t_1^{(2)}$. We are trying to infer whether the second patient's TB disease is latent reactivation or recent transmission. We only need to compute the probability in two cases.

i. $\overline{\mathcal{S}^{(2)}}$ The probability that we observe $t_1^{(2)}$ given $I^{(2)}$ came in with latent infection at $t_0^{(2)}$ and $I^{(1)}$ has entering/diagnosis times $t_0^{(1)}$,$t_1^{(1)}$ is

$$\begin{aligned} p(t_1^{(2)} | \overline{\mathcal{S}^{(2)}}, t_0^{(1)}, t_1^{(1)}, t_0^{(2)}) &= p(t_1^{(2)} | \overline{\mathcal{S}^{(2)}}, t_0^{(2)}) \\ &= \frac{\alpha\gamma[(1-\alpha)^{t_1^{(2)}-t_0^{(2)}+1} - ((1-\gamma)^{t_1^{(2)}-t_0^{(2)}+1})]}{\gamma - \alpha} \end{aligned} \tag{3.21}$$

Explanation of equation (3.21) Each month, there is a probability α for $I^{(2)}$'s latent infection to become active. After becoming active, he/she will have a chance of γ per month of being diagnosed. The length of time from $I^{(2)}$ entering to getting diagnosed is modeled by a sum of two geometric random variables with success probabilities α and γ respectively. The value is computed by equation (3.5).

ii. $\mathcal{S}^{(2)}$ $I^{(2)}$ can be infected by $I^{(1)}$ only if $I^{(1)}$ is active. The probability of $I^{(2)}$ being infected by $I^{(1)}$ in month k will depends on the probability that $I^{(1)}$ is active

in that month, therefore it will be a function of time, k.

$$\hat{\beta}(k) = \beta \cdot p_a(t_0^{(1)}, t_1^{(1)}, k) \tag{3.22}$$

where $p_a(t_0^{(1)}, t_1^{(1)}, k)$ is defined as in equation (3.20). Once $I^{(2)}$ is infected, the dynamics will be the same as in the 2-body method. After the infection, $I^{(2)}$ could enter one of the two routes: fast (with probability δ) or slow (with probability $1 - \delta$). In the fast route, $I^{(2)}$ becomes active immediately after the infection and remains active with a probability of γ being diagnosed every month. In the slow route, $I^{(2)}$ first acquires latent infection and with probability α per month becomes active. Once $I^{(2)}$'s infection was activated, there is probability γ for he/she getting diagnosed every month. This case is illustrated in Figure 3.7.

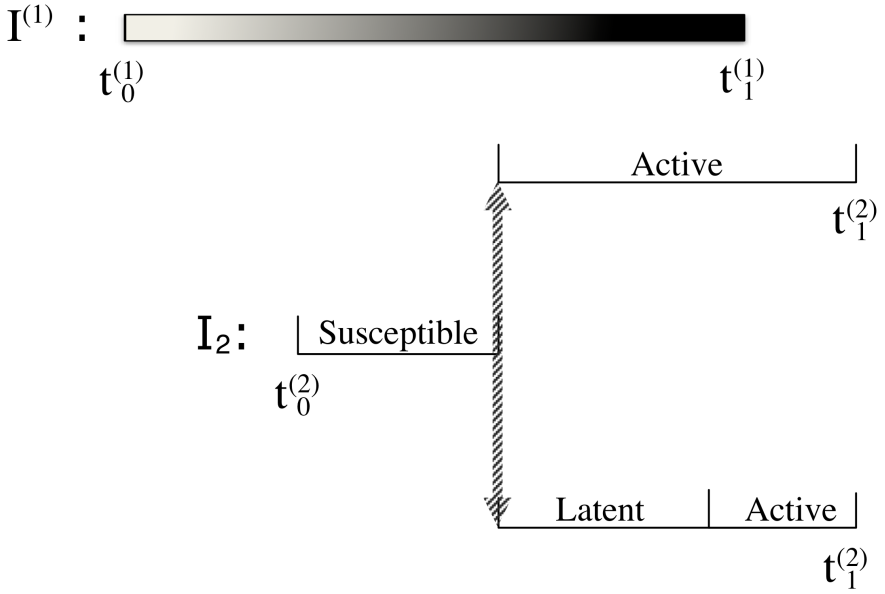


Figure 3.7: The likelihood of $I^{(1)}$ being infectious is represented by the color of the first bar: the deeper the color the more likely that $I^{(1)}$ is infectious at that time. If the month falls out of the range of $[t_0^{(1)}, t_1^{(1)}]$, the likelihood is 0. The probability that $I^{(2)}$ will be infected by $I^{(1)}$ is a function of time, $\hat{\beta}(k)$. Once $I^{(2)}$ is infected, the dynamics will be the same as in the 2-body method.

The conditional probability, given $I^{(2)}$ is susceptible at the time of entering,

and we observed $t_0^{(1)}$, $t_0^{(2)}$ and $t_1^{(1)}$ will be

$$\begin{aligned} p(t_1^{(2)}|\mathcal{S}^{(2)}, t_0^{(1)}, t_1^{(1)}, t_0^{(2)}) &= \sum_{j=t_0^{(2)}}^{t_1^{(2)}} \prod_{k=t_0^{(2)}}^{j-1} (1-\hat{\beta}(k))\hat{\beta}(j) \cdot \big[\delta(1-\gamma)^{t_1^{(2)}-j}\gamma \\ &+ (1-\delta)\frac{\alpha\gamma[(1-\alpha)^{t_1^{(2)}-j+1} - ((1-\gamma)^{t_1^{(2)}-j+1})]}{\gamma-\alpha}\big] \end{aligned} \quad (3.23)$$

Explanation of Equation (3.23): We are given that $I^{(2)}$ does not have latent infection at the time of entering the county and $I^{(1)}$ enters at $t_0^{(1)}$ and is diagnosed with TB disease at $t_0^{(2)}$. The chance that $I^{(2)}$ could be infected by $I^{(1)}$ will be a function of time, $\hat{\beta}(k)$ computed as in equation (3.20). $\prod_{k=t_0^{(2)}}^{j-1} (1-\hat{\beta}(k))\hat{\beta}(j)$ is the probability that $I^{(2)}$ is infected at month i. After the infection, $I^{(2)}$ has probability δ of entering the fast route and $1-\delta$ of entering the slow route. As in the 2-body method, in the fast route, the time from the infection to diagnosis will be modeled as the sum of two random variables with success probabilities α and γ. In the slow route, the time $I^{(2)}$ remains active before getting diagnosed will be modeled as a random variable with success probability γ.

Estimation of $p(\overline{\mathcal{S}^{(2)}}|\boldsymbol{t}^{(1)}, \boldsymbol{t}^{(2)})$ Now we have two conditional probabilities: $p(t_1^{(2)}|\mathcal{S}^{(2)}, t_0^{(1)}, t_1^{(1)}, t_0^{(2)})$ and $p(t_1^{(2)}|\overline{\mathcal{S}^{(2)}}, t_0^{(1)}, t_1^{(1)}, t_0^{(2)})$. Assume there are N_L people with latent infection and N_S susceptible. We assume the initial disease status of one patient is independent of his/her own entry time and the entry and diagnosis times of the other patient in the cluster. For convenience, we have the following definitions,

- $P_L = p(t_1^{(2)}, \overline{\mathcal{S}^{(2)}}|t_0^{(1)}, t_1^{(1)}, t_0^{(2)}) = p(t_1^{(2)}|t_0^{(1)}, t_1^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(2)}})\pi$
- $P_S = p(t_1^{(2)}, \mathcal{S}^{(2)}|t_0^{(1)}, t_1^{(1)}, t_0^{(2)}) = p(t_1^{(2)}|t_0^{(1)}, t_1^{(1)}, t_0^{(2)})(1-\pi)$

Following the same logic as in the **2-body method**, the conditional probability of observing $t_1^{(2)}$ and $I^{(2)}$'s initial status given the $t_0^{(1)}$, $t_1^{(1)}$, $t_0^{(2)}$ can be computed

as the following

$$\tilde{p}(t_1^{(2)}, \overline{\mathcal{S}^{(2)}} | t_0^{(1)}, t_1^{(1)}, t_0^{(2)}) = P_L H_{X_L}^{N_L - 1} H_{X_S}^{N_S} \tag{3.24}$$

$$\tilde{p}(t_1^{(2)}, \mathcal{S}^{(2)} | t_0^{(1)}, t_1^{(1)}, t_0^{(2)}) = P_S H_{X_L}^{N_L} H_{X_S}^{N_S - 1} \tag{3.25}$$

where H_{X_L} and H_{X_S} follows the same definition, i.e., equation (3.15) and (3.16). Finally, the conditional probabilities that we are interested in are computed as the following:

$$\begin{aligned} \tilde{p}(\overline{\mathcal{S}^{(2)}} | t_0^{(1)}, t_0^{(2)}, t_1^{(1)}, t_1^{(2)}) & \\ &= \frac{P_L H_{X_L}^{N_L - 1} H_{X_S}^{N_S}}{P_L H_{X_L}^{N_L - 1} H_{X_S}^{N_S} + P_S H_{X_L}^{N_L} H_{X_S}^{N_S - 1}} \\ &= \frac{P_L H_{X_S}}{P_L H_{X_S} + P_S H_{X_L}} \end{aligned} \tag{3.26}$$

$$\begin{aligned} \tilde{p}(\mathcal{S}^{(2)} | t_0^{(1)}, t_0^{(2)}, t_1^{(1)}, t_1^{(2)}) &= 1 - \tilde{p}(\overline{\mathcal{S}^{(2)}} | t_0^{(1)}, t_0^{(2)}, t_1^{(1)}, t_1^{(2)}) \\ &= \frac{P_S H_{X_L}}{P_L H_{X_S} + P_S H_{X_L}} \end{aligned} \tag{3.27}$$

3.3.3 The 1-body Mean Field Method for n-person Case

Now let's consider the case with n foreign-born patients, $I^{(1)}$, $I^{(2)}$,..., $I^{(n)}$. Among these n patients, assume $I^{(a)}$ is the patient whose latent status we are interested in knowing. We need to compute conditional probabilities in the following two cases.

i. $\overline{\mathcal{S}^{(a)}}$ Given that $I^{(a)}$ has latent infection at the time of entering the country, then $I^{(a)}$ will be independent of $I^{(b)}, b = 1, 2, ..., n, b \neq a$. The conditional probability of observing $t_1^{(a)}$ given $\overline{\mathcal{S}^{(a)}}$, $\{\boldsymbol{t}^{(b)}\}_{b=1...n, b \neq a}$ and $t_0^{(a)}$ can be computed with the same method as in the two person case, equation (3.21).

$$
\begin{aligned}
& p(t_1^{(a)}|\overline{\mathcal{S}^{(a)}}, \{\boldsymbol{t}^{(b)}\}_{b=1\ldots n, b\neq a}, t_0^{(a)}) = p(t_1^{(a)}|\overline{\mathcal{S}^{(a)}}, t_0^{(a)}) \\
= \quad & \frac{\alpha\gamma[(1-\alpha)^{t_1^{(a)}-t_0^{(a)}+1} - ((1-\gamma)^{t_1^{(a)}-t_0^{(a)}+1})]}{\gamma-\alpha}
\end{aligned} \tag{3.28}
$$

ii. $\mathcal{S}^{(a)}$ Given that $I^{(a)}$ enters the country without latent infection, he/she will be infected by one of the patients in $\{I^{(b)}\}_{i=1\ldots n, i\neq j}$. As in the two person case, the likelihood that patient $I^{(b)}$ is infectious in month k is assumed to be geometrically decay according to the length of time from k to $t_1^{(b)}$, with success probability γ. This likelihood, $p_a(t_0^{(b)}, t_1^{(b)}, k)$, can be computed using the same equation (3.20). The chance that $I^{(a)}$ is infected by $I^{(b)}$ at month k will be a function of k,

$$
\hat{\beta}_i(k) = \beta \cdot p_a(t_0^{(b)}, t_1^{(b)}, k) \tag{3.29}
$$

We denote the probability being infected by $I^{(b)}$ by $\beta_i(k)$. Since $I^{(a)}$ could be infected by any one of $\{I^{(b)}\}_{b=1\ldots n, b\neq a}$, the chance of being infected by each of the patients need to be combined to create a single infection probability. The probability that $I^{(a)}$ is infected at a given month k is,

$$
\tilde{\beta}(k) = 1 - \prod_{b=1, b\neq a}^{n} [1 - \hat{\beta}_i(k)] \tag{3.30}
$$

where $\hat{\beta}_i(k)$ is the defined in equation (3.29).

For example, suppose there are 4 individuals in a TB cluster. We would like to investigate the infectivity contributed by the first 3 patients. The entry and diagnosis times of the first three patients are the following: $I^{(1)} : [1, 16]$, $I^{(2)} : [6, 18]$ and $I^{(3)} : [7, 25]$. Assume $\beta = 0.5$ and $\gamma = 0.3$, the probability being infected by each patient and by all three are shown in Figure 3.8. The probability being infected by the b^{th} patient will be at its maximum value in the month of diagnosis. It decays geometrically when going towards the entry time. The probability being infected by one particular patient is zero in the months outside range he/she spend in the country, i.e. from entry to diagnosis.

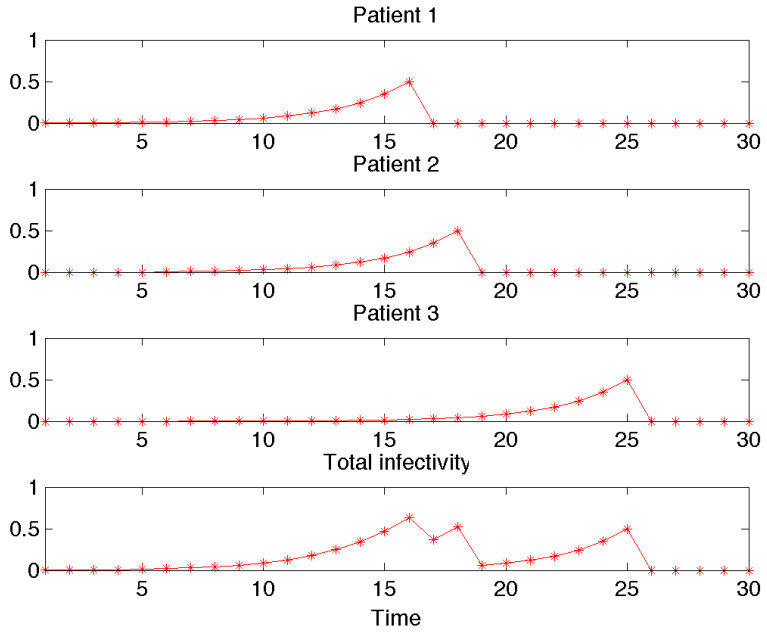


Figure 3.8: The entry and diagnosis times of the three individuals are $I^{(1)} : [1, 16]$, $I^{(2)} : [6, 18]$ and $I^{(3)} : [7, 25]$. Assume $\beta = 0.5$ and $\gamma = 0.3$. The infectivities contributed by each individual will have a maximum value of 0.5 at the diagnosis time and geometrically decay, with success probability 0.3, going further towards entry time and away from the diagnosis time. The probability being infected by one particular patient is zero in the month outside the range of time from his/her entry to diagnosis. The probability being infected by any of the three patients will be a combination of the three.

An illustration of the 1-body mean field method is shown in Figure 3.9. The periods of time $\{I^{(b)}\}_{b=1,2,\ldots,n,b\neq a}$ spend between entering the country to being diagnosed are presented by bars. The depth of the color indicates the likelihood that $I^{(b)}$ is infectious at that time: the darker the color the more likely. The abilities of $I^{(b)}$ to infect $I^{(a)}$ are combined to create a "super-individual", represented by the multi-sectional bar. The probability that the "super-individual" infects $I^{(a)}$ at month k is $\tilde{\beta}(k)$, as defined in equation (3.30). For the case of $I^{(a)}$, the dynamics are the same as in the two person 1-body mean field method.

Once we have the probability of $I^{(a)}$ being infected at each month k, we can compute the conditional probability of observing $t_1^{(a)}$ given $I^{(a)}$ came in without

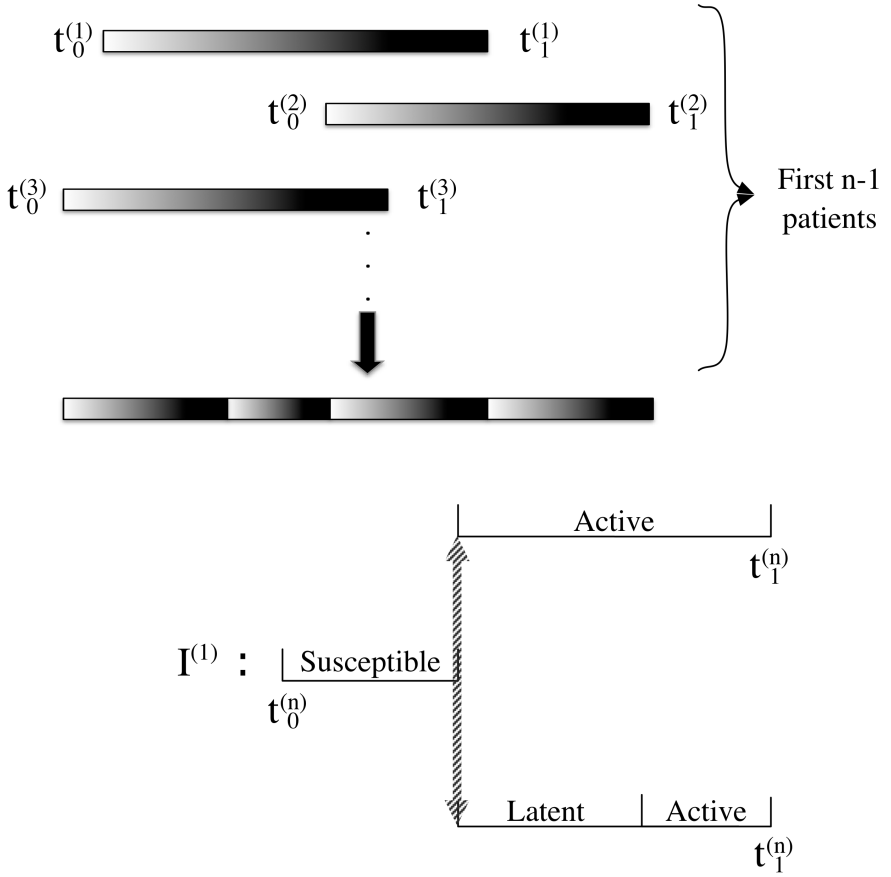


Figure 3.9: The likelihood of $\{I^{(b)}\}_{b=1...n,b\neq a}$ being infectious is represented by the color depth of the first n-1 bars: the deeper the color the more likely that $I^{(b)}$ is infectious at that time. If the month falls out of the range of $[t_0^{(b)}, t_1^{(b)}]$, the likelihood is 0. $I^{(a)}$ could be infected by any one of $\{I^{(b)}\}_{i=1...n,i\neq j}$, who are combined in into a "super-individual". Each month k the probability that $I^{(a)}$ is infected by this "super-individual" is $\tilde{\beta}(k)$, as defined in equation (3.30). The value of $\tilde{\beta}(k)$ is represented by the depth of the color of multi-sectional bar. Once $I^{(a)}$ is infected, the dynamics will be the same as in the 2-body method.

latent infection , $\boldsymbol{t}^{(b)}, b = 1, 2, ..., n, b \neq a$ and $t_0^{(a)}$. The computation follows the same formula as the two person case, i.e. equation (3.23).

$$p(t_1^{(a)}|\mathcal{S}^{(a)},\{\boldsymbol{t}^{(b)}\}_{i=1\ldots n,i\neq j},t_0^{(a)}) = \sum_{i=t_0^{(a)}}^{t_1^{(a)}}\prod_{k=t_1^{(a)}}^{i-1}(1-\tilde{\beta}(k))\tilde{\beta}(i)\cdot\big[\delta(1-\gamma)^{t_1^{(a)}-i}\gamma + (1-\delta)\frac{\alpha\gamma[(1-\alpha)^{t_1^{(a)}-i+1}-((1-\gamma)^{t_1^{(a)}-i+1})]}{\gamma-\alpha}\big] \tag{3.31}$$

Other than the n patients in the cluster, we assume there are N_L people with latent infection and N_S susceptible. Let us denote $p(t_1^{(a)}|\mathcal{S}^{(a)},\{\boldsymbol{t}^{(b)}\}_{b=1,2,\ldots,n,b\neq a},t_0^{(a)})$ as P_S and $p(t_1^{(a)}|\overline{\mathcal{S}^{(a)}},\{\boldsymbol{t}^{(b)}\}_{i=1,2,\ldots,n,b\neq a},t_0^{(a)})$ as P_L. The conditional probabilities $\tilde{p}(t_1^{(a)},\mathcal{S}^{(a)}|\{\boldsymbol{t}^{(b)}\}_{b=1,2,\ldots,n,b\neq a},t_0^{(a)})$ and $\tilde{p}(t_1^{(a)},\overline{\mathcal{S}^{(a)}}|\{\boldsymbol{t}^{(b)}\}_{b=1,2,\ldots,n,b\neq a},t_0^{(a)})$ can be computed using equation (3.26) and (3.27). The conditional probability $\tilde{p}(\overline{\mathcal{S}^{(a)}}|\boldsymbol{t}^{(1)},\ldots,\boldsymbol{t}^{(n)})$ can be computed as follows,

$$\tilde{p}(\overline{\mathcal{S}^{(a)}}|\boldsymbol{t}^{(1)},\ldots,\boldsymbol{t}^{(n)}) = \tag{3.32}$$

$$\frac{\tilde{p}(t_1^{(a)},\mathcal{S}^{(a)}|\{\boldsymbol{t}^{(b)}\}_{b=1,2,\ldots,n,b\neq a},t_0^{(a)})}{\tilde{p}(t_1^{(a)},\mathcal{S}^{(a)}|\{\boldsymbol{t}^{(b)}\}_{b=1,2,\ldots,n,b\neq a},t_0^{(a)})+\tilde{p}(t_1^{(a)},\overline{\mathcal{S}^{(a)}}|\{\boldsymbol{t}^{(b)}\}_{b=1,2,\ldots,n,b\neq a},t_0^{(a)})} \tag{3.33}$$

3.4 Infection Bath

In the previous section, we ignore the infectivity contributed by the foreign-born TB patients who have not yet been diagnosed and the ones in the domestic population. To make our model more realistic, we need to include the background infection bath. Assume there are hidden active patients, who have the same TB strain with our foreign-born patient cluster, in the foreign and domestic population. We see these patients as one super-individual with constant infectious rate. Each month, each individual in the patient's cluster will have a probability β_s of being infected by this super individual.

3.4.1 2-body Method

For a cluster with 2 patients, unlike the previous case, both patients could enter susceptible. This is because the TB bacteria could enter the cluster from the super-individual (the infection bath). Therefore there will be four different cases and we need to compute the probability of observing the entry and diagnosis times given all these cases.

i. $\overline{\mathcal{S}^{(1)}}\ \overline{\mathcal{S}^{(2)}}$ If both of the patients enter with latent infection, the super-individual has no impact on $I^{(1)}$ and $I^{(2)}$. This case is exactly the same as the model without the domestic infection bath. The computation for the probability of observing $\boldsymbol{t}^{(1)}$ and $\boldsymbol{t}^{(2)}$ and $\overline{\mathcal{S}^{(1)}}\ \overline{\mathcal{S}^{(2)}}$ follows equations (3.6).

ii. $\overline{\mathcal{S}^{(1)}}\mathcal{S}^{(2)}$ In this case, $I^{(1)}$ enters with latent infection while $I^{(2)}$ enters susceptible. Before $I^{(1)}$ becomes active or after he/she is diagnosed, $I^{(2)}$ has the risk to be infected by the super-individual with the probability β_s each month. In the period while $I^{(1)}$ is active, the collective transmission probability of $I^{(1)}$ and the super-individual is β_t per month, where $\beta_t = 1 - (1-\beta)(1-\beta_s)$. Figure 3.10 shows an illustration of the transmission dynamics of this case.

Given $I^{(1)}$ has latent infection, $I^{(2)}$ is susceptible at the entry time and $I^{(1)}$ becomes active in the i^{th} month, $t_0^{(1)} \leq i \leq t_1^{(1)}$, we would like to compute the conditional probability that we observed diagnosis time:

$$p\left(t_1^{(2)} | \boldsymbol{t}^{(1)}, t_0^{(2)}, I^{(1)}\text{active at } i, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)}\right)$$

. The computation is different from the model without the background infection rate since the probability being infected by $I^{(1)}$ in a specific month depends on whether $I^{(1)}$ is active in that month. Let's denote the probability being infected by $I^{(1)}$ in month t, given $I^{(1)}$ becomes active in month i and is diagnosed at $t_1^{(1)}$ as $\beta(i, t_1^{(1)}, t)$. It is defined as follows:

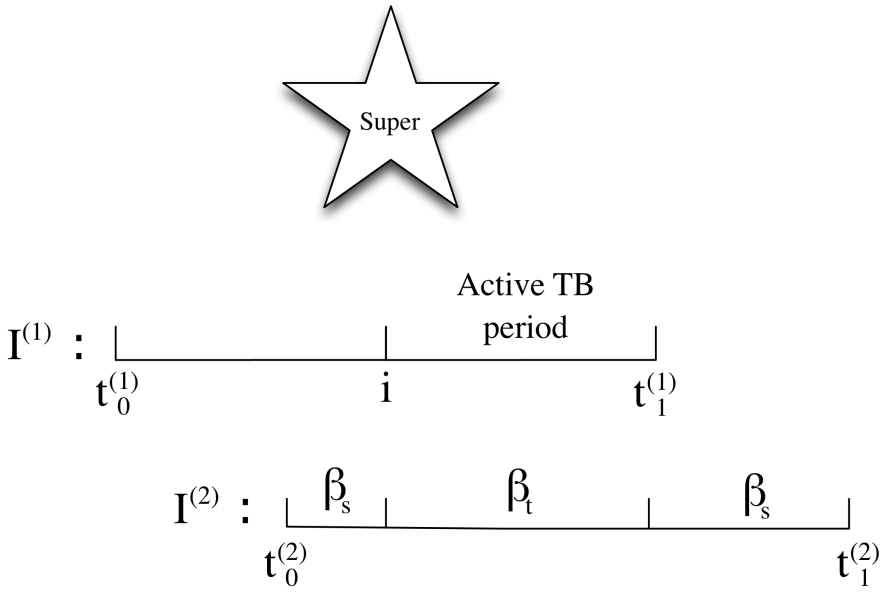


Figure 3.10: An illustration of the 2-person case with a domestic infection bath represented as a super individual with constant transmission rate. Assume $I^{(1)}$ is active from i to $t_1^{(1)}$, within this period, $I^{(2)}$ has a probability β_t being infected. Other than the period from i to $t_1^{(1)}$, $I^{(2)}$ has a probability β_s being infected.

$$\beta(i, t_1^{(1)}, t) = \begin{cases} 1 - (1-\beta)(1-\beta_s) & \text{if } i \le t \le t_1^{(1)} \\ \beta_s & \text{otherwise} \end{cases} \tag{3.34}$$

In the period when $I^{(1)}$ is active, the probability for $I^{(2)}$ being infected is a result bases on the combination of the infectivity from $I^{(1)}$ and the background rate. In the period when $I^{(1)}$ has not yet become active or he/she is diagnosed, the only infecitivity for $I^{(2)}$ comes from the background rate. The conditional probability is computed as the following:

$$
\begin{aligned}
& p\left(t_1^{(2)} | t_0^{(1)}, t_1^{(1)}, t_0^{(2)}, I^{(1)}\text{active at } i, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)}\right) \\
= & \sum_{j=t_0^{(2)}}^{t_1^{(2)}} \left[\prod_{k=t_0^{(1)}}^{j-1} (1-\beta(i, t_1^{(1)}, k))\right] \beta(i, t_1^{(1)}, j) f(j, t_1^{(2)})
\end{aligned} \tag{3.35}
$$

where $f(j,t)$ computes the probability that a susceptible patient in the cluster, which is $I^{(2)}$ in the case, that was infected in month j will be diagnosed in month t.

$$
f(j,t) = \delta(1-\gamma)^{t-j}\gamma + (1-\delta)\frac{\alpha\gamma[(1-\alpha)^{t-j+1} - (1-\gamma)^{t-j+1}]}{\gamma - \alpha} \tag{3.36}
$$

Explanation of Equation (3.35): Suppose $I^{(2)}$ is infected in month j, then for month $t_0^{(2)}$ to $j-1$, the probability that $I^{(2)}$ is not infected is $1-\beta(i, t_1^{(1)}, k)$. The probability that $I^{(2)}$ is infected in month j is $\beta(i, t_1^{(1)}, j)$. This gives us the first part in equation (3.35): $\left[\prod_{k=t_0^{(1)}}^{j-1}(1-\beta(i, t_1^{(1)}, k))\right]\beta(i, t_1^{(1)}, j)$. Once $I^{(2)}$ is infected in month j, the dynamics is the same as that of model in the previous section, i.e. the model without the background infectivity. The computation for the probability that $I^{(2)}$ will be diagnosed in $t_1^{(2)}$, given he/she was infected in month j, is expressed in equation (3.36).

Given $I^{(1)}$ has latent infection at the time of entry, the event "$I^{(1)}$ becomes active in month i " is independent of the initial status and the entry time of $I^{(2)}$. Therefore we have the following expression

$$
\begin{aligned}
& p\left(I^{(1)}\text{active at } i, t_1^{(1)} | t_0^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)}\right) \\
= & \; p\left(I^{(1)}\text{active at } i, t_1^{(1)} | t_0^{(1)}, \overline{\mathcal{S}^{(1)}}\right) \\
= & \; (1-\alpha)^{i-t_0^{(1)}}\alpha(1-\gamma)^{t_1^{(1)}-i}\gamma
\end{aligned} \tag{3.37}
$$

The conditional probability of observing $t_1^{(1)}$ and $t_1^{(2)}$ given $t_0^{(1)}$, $t_0^{(2)}$, $\overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)}$ could be computed by looping over all the possible values for the time when $I^{(1)}$ becomes

active.

$$
\begin{aligned}
& p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)}) \\
= & \sum_{i=t_0^{(1)}}^{t_1^{(1)}} \left[p\left(I^{(1)}\text{active at } i, t_1^{(1)} | t_0^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)} \right) \right. \qquad (3.38) \\
& \left. \cdot p\left(t_1^{(2)} | t_0^{(1)}, t_1^{(1)}, t_0^{(2)}, I^{(1)}\text{active at } i, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)} \right) \right] \\
= & \sum_{i=t_0^{(1)}}^{t_1^{(1)}} \left[(1-\alpha)^{i-t_0^{(1)}} \alpha (1-\gamma)^{t_1^{(1)}-i} \gamma \right. \qquad (3.39) \\
& \left. \cdot p\left(t_1^{(2)} | t_0^{(1)}, t_1^{(1)}, t_0^{(2)}, I^{(1)}\text{active at } i, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)} \right) \right] \qquad (3.40)
\end{aligned}
$$

where $p\left(t_1^{(2)} | t_0^{(1)}, t_1^{(1)}, t_0^{(2)}, I^{(1)}\text{active at } i, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)} \right)$ is computed as in equation (3.36)

iii. $\mathcal{S}^{(1)}$ $\overline{\mathcal{S}^{(2)}}$ This case will be the same as $\overline{\mathcal{S}^{(1)}}\mathcal{S}^{(2)}$, after exchanging the patient index.

iv. $\mathcal{S}^{(1)}$ $\mathcal{S}^{(2)}$ The scenario in which $I^{(1)}$ and $I^{(2)}$ both enter without latent infections is possible now, since the TB bacteria could enter the cluster from the domestic infection bath. If the two periods that $I^{(1)}$ and $I^{(2)}$ spend from entry to diagnosis do not overlap, i.e. $t_1^{(1)} < t_0^{(2)}$ or $t_1^{(2)} < t_0^{(1)}$, the only possible scenario will be both $I^{(1)}$ and $I^{(2)}$ are infected by the super-individual. The conditional probability of this scenario will be

$$
\begin{aligned}
p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \mathcal{S}^{(1)}, \mathcal{S}^{(2)}) = & \sum_{i=t_0^{(1)}}^{t_1^{(1)}} (1-\beta_s)^{i-t_0^{(1)}} \beta_s f(i, t_1^{(1)}) \\
& \cdot \sum_{i=t_0^{(2)}}^{t_1^{(2)}} (1-\beta_s)^{i-t_0^{(2)}} \beta_s f(i, t_1^{(2)}) \qquad (3.41)
\end{aligned}
$$

where $f(i,t)$ is defined as in equation (3.36). If the periods overlap, the computation depends on four time points: 1. time when $I^{(1)}$ is infected (τ_1); 2. time when $I^{(1)}$ becomes active (τ_2); 3. time when $I^{(2)}$ is infected (τ_3); and 4. time when $I^{(2)}$ becomes active (τ_4). We define $\beta_2(t)$ to be the probability for $I^{(1)}$ being infected by $I^{(2)}$. For example, in the month when $I^{(2)}$ is not active, $\beta_2(t) = \beta_s$ and $\beta_2(t) = 1-(1-\beta_s)(1-\beta)$ if $I^{(2)}$ is active. Similarly, we have the probability for $I^{(2)}$ being infected by $I^{(1)}$ as $\beta_1(t)$. The probability for $I^{(1)}$ being infected by $I^{(2)}$ in month t is $\beta(t,\tau_4,t_1^{(2)})$, which is defined as in equation (3.34). Similarly, the probability for $I^{(2)}$ being infected by $I^{(1)}$ in month t is $\beta(t,\tau_2,t_1^{(1)})$. Let us define function $h(\tau_1,\tau_2,\tau_3,\tau_4)$ as follows,

$$\begin{aligned} h(\tau_1,\tau_2,\tau_3,\tau_4) &= \prod_{k=t_0^{(1)}}^{\tau_1-1}[1-\beta(\tau_4,t_1^{(2)},k)]\beta(\tau_4,t_1^{(2)},\tau_1)g(\tau_1,\tau_2,t_1^{(1)}) \\ &\quad \cdot \prod_{k=t_0^{(2)}}^{\tau_3-1}[1-\beta(\tau_2,t_1^{(1)},k)]\beta(\tau_2,t_1^{(1)},\tau_3)g(\tau_3,\tau_4,t_1^{(2)}) \end{aligned} \tag{3.42}$$

where $g(a,b,t)$ is a helper function defined as follows,

$$g(a,b,t) = \mathcal{I}_{\{a=b\}}\delta(1-\gamma)^{t-a}\gamma + (1-\delta)(1-\alpha)^{b-a}\alpha(1-\gamma)^{t-b}\gamma \tag{3.43}$$

where $\mathcal{I}_{\{a=b\}}$ is an indicator function with value 1 if $a=b$ (0 otherwise). The function $g(a,b,t)$ in equation (3.43) computes the probability that a patient is infected in month a and is diagnosed in month t. There is a probability δ that this patient becomes active immediately after infection. There is a probability $1-\delta$ that the patient becomes latent and becomes active later in month b. The final conditional probability can be computed by looping over all possible values for τ_1-τ_4,

$$p(t_1^{(1)},t_1^{(2)}|t_0^{(1)},t_0^{(2)},\mathcal{S}^{(1)},\mathcal{S}^{(2)}) = \sum_{\tau_1=t_0^{(1)}}^{t_1^{(1)}}\sum_{\tau_2=\tau_1}^{t_1^{(1)}}\sum_{\tau_3=t_0^{(2)}}^{t_1^{(2)}}\sum_{\tau_4=\tau_3}^{t_1^{(2)}} h(\tau_1,\tau_2,\tau_3,\tau_4) \tag{3.44}$$

Since computing equation (3.44), which involves 4 loops, will be extremely expensive, we make approximations to simplify the transmission dynamics in this case.

We assume there are only three scenarios for $I^{(1)}$ and $I^{(2)}$ to be infected: 1) $I^{(1)}$ is infected by the background bath first. $I^{(2)}$ is infected by $I^{(1)}$ or the background later; 2) $I^{(2)}$ is infected by the background bath first. $I^{(1)}$ is infected by $I^{(2)}$ or the background later; 3) $I^{(1)}$ and $I^{(2)}$ are infected by the background at the same time. Note that we skipped the case that both patients are infected by the background at different time. This is because that this case is considered in case 1) and 2). For scenario 1, the probability is computed as follows:

$$p_1 = \sum_{j=t_0^{(1)}}^{\min(t_1^{(1)},t_1^{(2)})} \left[\sum_{i=t_0^{(1)}}^{j} (1-\beta_s)^{(i-t_0^{(1)})} \beta_s g(i,j,t_1^{(1)}) \sum_{k=j}^{\min(t_1^{(1)},t_1^{(2)})} (1-\tilde{\beta})^{(k-j)} \tilde{\beta} f(k,t_1^{(2)}) \right] \tag{3.45}$$

where $g(i,j,t_1^{(1)})$ is defined as in equation (3.43). $\sum_{i=t_0^{(1)}}^{j} (1-\beta_s)^{(i-t_0^{(1)})} \beta_s g(i,j,t_1^{(1)})$ computes the probability of $I^{(1)}$ being infected by the background (in any month from $t_0^{(1)}$ to j), becoming active in month j and is diagnosed at $t_1^{(1)}$. $f(k,t_1^{(2)})$ is defined as in equation (3.36) and $\sum_{k=j}^{\min(t_1^{(1)},t_1^{(2)})} (1-\tilde{\beta})^{(k-j)} \tilde{\beta} f(k,t_1^{(2)})$ computes the probability $I^{(2)}$ is infected by $I^{(1)}$ or the background (after $I^{(1)}$ becomes active and before being diagnosed) and is diagnosed at $t_1^{(2)}$. $\tilde{\beta}$ is the combined infectivity of $I^{(1)}$ and the background. It is defined as $\tilde{\beta} = 1-(1-\beta)(1-\beta_s)$. For scenario 2, the computation will be the same as scenario one after switching the patient index. The probability of this scenario is denoted as p_2.

For scenario 3, the probability p_3 is computed as follows,

$$\begin{aligned} p_3 &= \sum_{i=\tau_a}^{t_1^{(1)}} (1-\beta_s)^{(i-t_0^{(1)})} \beta_s f(i,t_1^{(2)}) (1-\beta_s)^{(i-t_0^{(2)})} \beta_s f(i,t_1^{(2)}) \\ &= \sum_{i=\tau_a}^{t_1^{(1)}} (1-\beta_s)^{(2i-t_0^{(1)}-t_0^{(2)})} \beta_s^2 f(i,t_1^{(2)}) f(i,t_1^{(2)}) \end{aligned} \tag{3.46}$$

where $\tau_a = \max(t_0^{(1)}, t_0^{(2)})$. Finally, $p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \mathcal{S}^{(1)}, \mathcal{S}^{(2)}) = p_1 + p_2 + p_3$. We now have computed:

1. $p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \mathcal{S}^{(1)}, \mathcal{S}^{(2)})$
2. $p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(1)}}, \mathcal{S}^{(2)})$
3. $p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \mathcal{S}^{(1)}, \overline{\mathcal{S}^{(2)}})$
4. $p(t_1^{(1)}, t_1^{(2)} | t_0^{(1)}, t_0^{(2)}, \overline{\mathcal{S}^{(1)}}, \overline{\mathcal{S}^{(2)}})$

With these four conditional probabilities ready, the ones that we are interested in, i.e. $p(\overline{\mathcal{S}^{(2)}} | t_0^{(1)}, t_1^{(1)}, t_0^{(2)}, t_1^{(2)})$ and $p(\mathcal{S}^{(2)} | t_0^{(1)}, t_1^{(1)}, t_0^{(2)}, t_1^{(2)})$ can be computed following the same procedure as in the **2-body method** without the background infection bath: equation (3.18) and (3.19).

In the event of clusters with more than two patients, we could compute with any two patients while treating them as a cluster of size 2 and putting everyone else in the background. For a cluster of size n, we choose the patients $I^{(a)}$ and $I^{(b)}$ and treat them as a cluster of size two. The probability being infected by the $n-2$ patients (other than $I^{(a)}$ and $I^{(b)}$) can be computed as in equation (3.30). This probability can be further combined with the constant infection probability β_s for the domestic background bath or β for the active patient within the cluster. For the target patient (either $I^{(a)}$ or $I^{(b)}$), the probability being infected by the background bath is now a function of time: $\hat{\beta}_s(t)$, which is defined as follows,

$$\hat{\beta}_s(t) = 1 - (1 - \tilde{\beta}(t))(1 - \beta_s) \tag{3.47}$$

where $\tilde{\beta}(t)$ is the probability of being infected by the patients $I^{(c)}$, $c = 1, 2, ..., n; c \neq a, b$, as in equation (3.30). The model is indifferent of choosing $I^{(a)}$ or $I^{(b)}$ as the target patient. For the purpose of illustration, let us choose $I^{(a)}$ as the target patient. The probability of $I^{(a)}$ being infected by $I^{(b)}$, given he/she is active in a month is β. The probability of $I^{(a)}$ being infected by the other patients (other than $I^{(b)}$) is $\tilde{\beta}(t)$. Therefore the probability being infected by the other patients within the cluster will be

$$\hat{\beta}(t) = 1 - (1 - \tilde{\beta}(t))(1 - \beta) \tag{3.48}$$

The transmission dynamics remain the same and we only need to modify background infection rate from a constant β_s to a function of time $\hat{\beta}_s(t)$. In equations (3.34), the constant probability β_s is now replaced by $\hat{\beta}_s(t)$. In equation (3.41) and (3.45), the part $(1-\beta_s)^{i-t_0^{(j)}}\beta_s$ $(j=1,2)$ computes the probability that $I^{(j)}$ enter susceptible at $t_0^{(j)}$ and is infected in month i. This part becomes $\prod_{k=t_0^{(j)}}^{i-1}(1-\hat{\beta}_s(k))\hat{\beta}_s(i)$, if $i > t_0^{(j)}$ or $\hat{\beta}_s(t_0^{(j)})$ if $i = t_0^{(j)}$. Similarly in equation (3.45) the part $(1-\beta)^{i-t_0^{(j)}}\beta$ $(j=1,2)$ is now: $\prod_{k=t_0^{(j)}}^{i-1}(1-\hat{\beta}(k))\hat{\beta}(i)$, if $i > t_0^{(j)}$ or $\hat{\beta}(t_0^{(j)})$ if $i = t_0^{(j)}$.

In this case, we are computing with two patients and one of them is the target patient. Since we put all the rest of the patients into the background and treat their infectivities in a similar manner as in the 1-body mean field method. Therefore, we also call the 2-body method as *2-body mean field method*.

3.4.2 1-body Mean Field Method

For the 1-body mean field method, adding the background infection bath is simple. We just need to include the probability of being infected by the domestic infection bath, in addition to the probability of being infected by other patients in the cluster, which is $\tilde{\beta}(t)$ in equation (3.30). The new probability of being infected by the background infection bath is computed using the same equation (3.47). The rest of the computation will be the same as the approximation model without the bath.

3.5 Conclusions

So far we have been built the theoretical background of our model. Based on the information we have, which is the entry and the diagnosis time of each patient in one TB cluster, we are able to estimate the conditional probability that any of them enter with latent infection (or susceptible). We have developed the 2-body method, which could compute the probabilities on a clusters of size 2. However, dud to the computational complexity of the 2-body method, the computation of the clusters of size larger than 2 is expensive. The 1-body mean field method solves this prob-

lem by simplifying the transmission dynamic. It largely reduces the computational complexity and be able to compute clusters with any size. Ideally, we will estimate the parameters from the data and apply our model to the data. The characteristics of the patients data we have prevent us from having an accurate estimation of the parameters. The next chapter will discuss the problem of parameter estimation through the analysis of Fisher Information.

CHAPTER 4
Parameter Estimation

4.1 Introduction

We have collected the TB patient data in New York city from November 2001 to December 2007 [46]. For each patient, we have the following information: 1) Spoligotype; 2) RFLP; 3) entry time; 4)diagnosis time and 5) country of birth. For each patient we have the DNA fingerprints which allows us to organize them into small clusters. All the patients in one TB cluster share the identical Spoligotype and RFLP.

Each patient has a finite life span. Therefore, every patient in our data was diagnosed with TB before he/she dies from natural causes. If the time a patient spent from entry to diagnosis is modeled by a random variable, then this random variable is truncated. Suppose there are n foreign-born persons in our model universe, our data could only contain those who are diagnosed before reaching the end of his/her life span. Since latent reactivation and transmission are rare events, only a small portion of n will be diagnosed. We argue that the data do not contain sufficient information to have an accurate estimation on the parameters we are interested in. We will illustrate this idea by using two simplified models.

We built two models with geometric random variables with truncation as analogies to the real data. Here we assume each individual in our model have the same life span: k. Individuals exist in our model for more than k months will be removed. The analysis of Fisher Information and Cramér-Rao Lower Bound shows that, for our simplified models, the variance of the estimated parameters increases as the truncation value k decreases. This is confirmed with results of the numerical experiments. In our TB model, where k is relatively small, the data we have is not enough for an accurate parameter estimation.

This chapter is organized in the following way. First, the simplified models are introduced. Next, data are simulated with various truncation values k and the parameters are estimated with Maximum Likelihood Estimation. Finally, an

analysis using Fisher Information is conducted to link theories to the simulation results and the available data sets.

4.2 Single GRV with Truncation

Given a GRV Y, with probability mass function

$$f(y|\alpha) = (1-\alpha)^y \alpha \qquad y = 0, 1, 2, ... \tag{4.1}$$

Now we consider the likelihood function of some truncated observations of Y, conditioned on $Y \leq k$, where k is a positive integer. Note that this is a simpler analogy of our model: Suppose all the people enter at the same time and enter with latent infection and each month a patient has probability α to be diagnosed. The life expectancy is k months. We observed those patients who are diagnosed before the k^{th} month.

Suppose $\{y_i\}_{i=1}^m$ are the samples from the geometric distribution with success probability α. Let $\{x_i\}_{i=1}^n$ be chosen from $\{y_i\}_{i=1}^m$ with the condition $x_i \leq k$. Therefore, the truncated GRV with the success probability α and truncation value k has the probability mass function as the following

$$f(x|\alpha) = \frac{(1-\alpha)^x \alpha}{1-(1-\alpha)^{k+1}} \qquad x = 0, 1, 2, ..., k \tag{4.2}$$

The likelihood function of the n observations with parameter α is

$$\mathcal{L}(\alpha) = \frac{(1-\alpha)^{\sum_{i=1}^n x_i} \alpha^n}{[1-(1-\alpha)^{k+1}]^n} \tag{4.3}$$

The log-likelihood will be

$$\log\mathcal{L}(\alpha) = \log(1-\alpha) \sum_{i=1}^{n} x_i + n\log(\alpha) - n\log[1-(1-\alpha)^{k+1}] \tag{4.4}$$

Taking the derivative of the log-likelihood respect to α, we have,

$$\frac{d\log\mathcal{L}(\alpha)}{d\alpha} = -\frac{\sum_{i=1}^n x_i}{1-\alpha} + \frac{n}{\alpha} - \frac{n(k+1)(1-\alpha)^k}{1-(1-\alpha)^{k+1}} \tag{4.5}$$

4.3 Hybrid GRV with Truncation

Let Y be a hybrid GRV such that with probability π, the success probability is α and with probability $1-\pi$, the success probability is β. The probability mass function of Y is

$$f(y|\pi) = \begin{cases} (1-\alpha)^y\alpha & \text{with probability } \pi \\ (1-\beta)^y\beta & \text{with probability } 1-\pi \end{cases} \tag{4.6}$$

Similarly, let $\{y_i\}_{i=1}^m$ be the samples from the pmf in equation (4.6). Let $\{x_i\}_{i=1}^n$ be chosen from $\{y_i\}_{i=1}^m$ with the condition $x_i \leq k$. The random variable, X, resulting from truncating Y at k, has the following probability mass function,

$$f(x|\pi) = \frac{(1-\beta)^x\beta + \pi[(1-\alpha)^x\alpha - (1-\beta)^x\beta]}{1-(1-\beta)^{k+1} - \pi[(1-\alpha)^{k+1} - (1-\beta)^{k+1}]} \quad x = 0, 1, 2, ..., k \tag{4.7}$$

The likelihood function of the observations with parameter π is

$$\mathcal{L}(\pi) = \prod_{i=1}^{n} \left\{ \frac{(1-\beta)^{x_i}\beta + \pi[(1-\alpha)^{x_i}\alpha - (1-\beta)^{x_i}\beta]}{1-(1-\beta)^{k+1} - \pi[(1-\alpha)^{k+1} - (1-\beta)^{k+1}]} \right\} \tag{4.8}$$

The log-likelihood is

$$\begin{aligned} \log\mathcal{L}(\pi) &= \sum_{i=1}^{n} \Big\{ \log\left((1-\beta)^{x_i}\beta + \pi[(1-\alpha)^{x_i}\alpha - (1-\beta)^{x_i}\beta]\right) \\ &\quad -\log(1-(1-\beta)^{k+1} - \pi[(1-\alpha)^{k+1} - (1-\beta)^{k+1}]) \Big\} \end{aligned} \tag{4.9}$$

Taking the derivative respect to π, we have

$$\frac{d\log\mathcal{L}(\pi)}{d\pi} = \sum_{i=1}^{n} \left\{ \frac{A(x_i)}{(1-\beta)^{x_i} + \pi A(x_i)} + \frac{B}{1-(1-\beta)^{k+1} - \pi B} \right\} \tag{4.10}$$

where $A(x_i) = (1-\alpha)^{x_i}\alpha - (1-\beta)^{x_i}\beta$ and $B = (1-\alpha)^{k+1} - (1-\beta)^{k+1}$. To find the π maximizes equation (4.9), we need to solve $\dfrac{d\log\mathcal{L}(\pi)}{d\pi} = 0$.

4.4 Asymptotic Behavior of MLE

Suppose we have n observations of random variable X, $\{x_i\}_{i=1}^n$. X follows the distribution whose density/mass function is $f(x|\theta)$, where θ is the parameter associated with the distribution. The log-likelihood will be

$$\log(\mathcal{L})(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{4.11}$$

Let θ^* is the Maximum Likelihood Estimator (MLE), i.e. θ^* maximizes $\log(\mathcal{L})(\theta)$. Supposing the true parameter is θ_0 and $n \to \infty$, we have the following [47],

$$(\theta^* - \theta_0) \sim \mathcal{N}\left(0, \frac{1}{nI(\theta_0)}\right) \tag{4.12}$$

Equation (4.12) states that the error of a maximum likelihood estimator is asymptotically normally distributed with mean 0 and variance $\frac{1}{nI(\theta_0)}$. $I(\theta_0)$ is the *Fisher Information*, which is defined as the following

$$I(\theta_0) = -\mathbb{E}\left[\frac{d^2 \log f(X|\theta)}{d\theta^2}|_{\theta=\theta_0}\right] \tag{4.13}$$

note that $\frac{1}{nI(\theta_0)}$ is also the Cramér-Rao Lower Bound (CRLB) for the variance of any unbiased estimator with n observations. Now we would like to compute the CRLB of the two estimators we showed before.

4.4.1 Single GRV with Truncation

Let X be a GRV with success probability α and truncation value k. Let us define function $f(X|\alpha)$ as the following

$$f(X|\alpha) = \frac{(1-\alpha)^X \alpha}{1-(1-\alpha)^{k+1}} \tag{4.14}$$

The expectation of X is

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x=0}^{k} x\frac{(1-\alpha)^x\alpha}{1-(1-\alpha)^{k+1}} \\
&= \frac{(1-\alpha)\alpha}{1-(1-\alpha)^{k+1}}\sum_{x=0}^{k} x(1-\alpha)^{x-1} \\
&= \frac{(1-\alpha)\alpha}{1-(1-\alpha)^{k+1}}\left[-\frac{d}{d\alpha}\sum_{x=0}^{k}(1-\alpha)^x\right] \\
&= \frac{(1-\alpha)\alpha}{1-(1-\alpha)^{k+1}}\left[-\frac{d}{d\alpha}\frac{1-(1-\alpha)^{k+1}}{\alpha}\right] \\
&= \frac{1-\alpha}{1-(1-\alpha)^{k+1}}\frac{1-(k+1)(1-\alpha)^k\alpha-(1-\alpha)^{k+1}}{\alpha}
\end{aligned} \tag{4.15}$$

The Fisher information of α is computed using the following steps.

$$\log f(X|\alpha) = X\log(1-\alpha) + \log(\alpha) - \log\left(1-(1-\alpha)^{k+1}\right) \tag{4.16}$$

$$\frac{d\log f(X|\alpha)}{d\alpha} = -\frac{X}{1-\alpha} + \frac{1}{\alpha} - \frac{(k+1)(1-\alpha)^k}{1-(1-\alpha)^{k+1}} \tag{4.17}$$

$$\begin{aligned}
\frac{d^2\log f(X|\alpha)}{d\alpha^2} &= -\frac{X}{(1-\alpha)^2} - \frac{1}{\alpha^2} \\
&\quad + \frac{(k+1)k(1-\alpha)^{k-1}[1-(1-\alpha)^{k+1}] + (k+1)^2(1-\alpha)^{2k}}{[1-(1-\alpha)^{k+1}]^2}
\end{aligned} \tag{4.18}$$

Let α_0 be the true parameter, the Fisher Information is computed as the following

$$\begin{aligned}
I(\alpha_0) &= -\mathbb{E}\left[\frac{d^2\log f(X|\alpha)}{d\alpha^2}\bigg|_{\alpha=\alpha_0}\right] \\
&= \frac{\mathbb{E}[X]}{(1-\alpha_0)^2} + \frac{1}{\alpha_0^2} \\
&\quad - \frac{(k+1)k(1-\alpha_0)^{k-1}[1-(1-\alpha_0)^{k+1}] + (k+1)^2(1-\alpha_0)^{2k}}{[1-(1-\alpha_0)^{k+1}]^2}
\end{aligned} \tag{4.19}$$

The standard deviation of the estimator α^* is computed based on the CRLB i.e. $\sigma = \sqrt{\frac{1}{nI(\alpha_0)}}$.

Asymptotic Analysis When $k\alpha_0$ is small, we can assume the following

$$1-(1-\alpha_0)^{k+1} = (k+1)\alpha_0 - \frac{1}{2}(k+1)k\alpha_0^2 + O(k^3\alpha_0^3) \tag{4.20}$$

Now let us substitute equation (4.20) in to the third term of equation (4.19) and analyze the expression separately. The first term in the numerator becomes the following

$$\begin{aligned}
& (k+1)k(1-\alpha_0)^{k-1}[1-(1-\alpha_0)^{k+1}] \\
= \quad & (k+1)k[1-(k-1)\alpha_0 + \frac{1}{2}(k-1)(k-2)\alpha_0^2 + O(k^3\alpha_0^3)] \cdot [(k+1)\alpha_0 \\
& -\frac{1}{2}(k+1)k\alpha_0^2 + O(k^3\alpha_0^3)] \\
= \quad & (k+1)k[(k+1)\alpha_0 - \frac{1}{2}(k+1)k\alpha_0^2 - (k-1)(k+1)\alpha_0^2 + O(k^3\alpha_0^3)] \\
= \quad & (k+1)^2[k\alpha_0 - \frac{1}{2}k^2\alpha_0^2 - k(k-1)\alpha_0^2 + O(k^3\alpha_0^3)]
\end{aligned} \tag{4.21}$$

The second term in the numerator becomes

$$\begin{aligned}
(k+1)^2(1-\alpha_0)^{2k} &= (k+1)^2[1-k\alpha_0 + \frac{1}{2}k(k-1)\alpha_0^2 + O(k^3\alpha_0^3)]^2 \\
&= (k+1)^2[1-2k\alpha_0 + k(k-1)\alpha_0^2 + k^2\alpha_0^2 + O(k^3\alpha_0^3)]
\end{aligned} \tag{4.22}$$

The numerator is now,

$$\begin{aligned}
& (k+1)k(1-\alpha_0)^{k-1}[1-(1-\alpha_0)^{k+1}] + (k+1)^2(1-\alpha_0)^{2k} \\
= \quad & (k+1)^2[1-k\alpha_0 + \frac{1}{2}k^2\alpha_0^2 + O(k^3\alpha_0^3)]
\end{aligned} \tag{4.23}$$

Apply Taylor expansion on the denominator, keep to $k^3\alpha_0^3$ and then use negative binomial expansion, we have the following

$$\begin{aligned} & [1-(1-\alpha_0)^{k+1}]^{-2} \\ = \; & [(k+1)\alpha_0 - \frac{1}{2}(k+1)k\alpha_0^2 + \frac{1}{6}(k+1)k(k-1)\alpha_0^3 + O(k^4\alpha_0^4)]^{-2} \\ = \; & (k+1)^{-2}\alpha_0^{-2}[1-(\frac{1}{2}k\alpha_0 - \frac{1}{6}k(k-1)\alpha_0^2 + O(k^3\alpha_0^3))]^{-2} \\ = \; & (k+1)^{-2}\alpha_0^{-2}[1+k\alpha_0 - \frac{1}{3}k(k-1)\alpha_0^2 + \frac{3}{4}k^2\alpha_0^2 + O(k^3\alpha_0^3)] \end{aligned} \tag{4.24}$$

Multiply the numerator term and the denominator term, we have

$$\begin{aligned} & \alpha_0^{-2}[1-k\alpha_0+\frac{1}{2}k^2\alpha_0^2+O(k^3\alpha_0^3)][1+k\alpha_0-\frac{1}{3}k(k-1)\alpha_0^2+\frac{3}{4}k^2\alpha_0^2+O(k^3\alpha_0^3)] \\ = \; & \frac{1}{\alpha_0^2} - \frac{k^2}{12} + \frac{k}{3} + O(k^3\alpha_0) \\ = \; & \frac{1}{\alpha_0^2} - \frac{k^2}{12} + O(k) + O(k^3\alpha_0) \end{aligned} \tag{4.25}$$

Since $X \leq k$, we have $\mathbb{E}[X] \approx O(k)$. Combining the three terms, we have the asymptotic expression of the Fisher Information for the single GRV with truncation,

$$\hat{I}(\alpha_0) \approx \frac{k^2}{12} + O(k) + O(k^3\alpha_0) \tag{4.26}$$

The exact and approximated values of Fisher Information of the truncated GRV are plot against different truncation values in Figure 4.1.

Comments The leading term of the Fisher Information of the single GRV with truncation is $\frac{k^2}{12}$. Let's say we have n data points and we would like to have an estimator of α with standard deviation $0.1\alpha_0$. We have the following,

$$\sqrt{\frac{12}{nk^2}} \;=\; 0.1\alpha_0 \tag{4.27}$$

Solving for n, we have

$$n = \frac{1200}{\alpha_0^2 k^2} \tag{4.28}$$

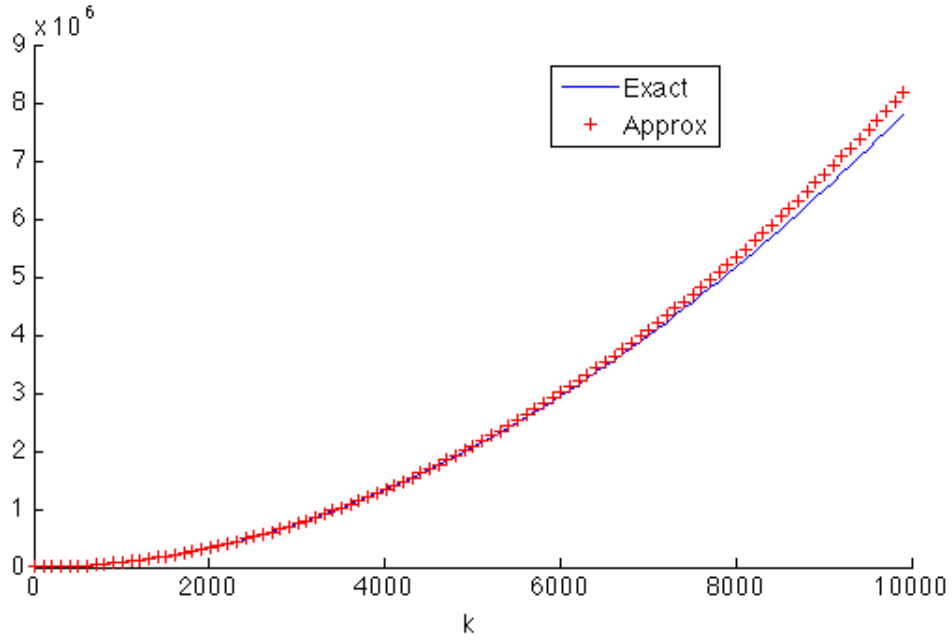


Figure 4.1: The approximation value of the Fisher Information as in equation (4.26) and the exact values as in equation (4.19) are plotted against different truncation value k.

If $\alpha_0 k = 0.05$, then we need 480000 data points in order to have an estimator for α with 10% standard deviation. Again under the condition of $\alpha_0 k = 0.05$, with only 5000 data points, the standard deviation of the estimator of α will be around $0.98\alpha_0$. This model is a simplified analogy to the case for foreign-born individuals with latent TB. Suppose the probability of becoming active given the individual is in latent state is α_0 per month. Since the time from becoming active to being diagnosed is relatively short comparing to the time from entry to becoming active, we ignore this time. We also assume each latent individual have a risk of 5% to develop active disease in the life time [6]. The time for a TB patient to spend from entry to diagnosis can be modeled as a GRV with success probability α_0 and truncation value k, such that $\alpha_0 k = 0.05$. We only have 3,741 data points and this includes the patients who were infected after entry. Therefore, we conclude that we don't have enough data to have an accurate estimator of α.

4.4.2 Hybrid GRV with Truncation

We would also like to compute the Fisher Information of the hybrid GRV with truncation. Let X be the the random variable. Let us define function $f(X|\pi)$ as the following

$$f(X|\pi) = \frac{(1-\beta)^X\beta + \pi[(1-\alpha)^X\alpha - (1-\beta)^X\beta]}{1-(1-\beta)^{k+1} - \pi[(1-\alpha)^{k+1} - (1-\beta)^{k+1}]} \tag{4.29}$$

For convenience, let us define $A(X) = (1-\alpha)^X\alpha - (1-\beta)^X\beta$ and $B = (1-\alpha)^{k+1} - (1-\beta)^{k+1}$. The first and second derivatives of $\log f(X|\pi)$ respect to π are computed as following

$$\frac{d\log f(X|\pi)}{d\pi} = \frac{A(X)}{(1-\beta)^X\beta + \pi A(X)} + \frac{B}{1-(1-\beta)^{k+1} - \pi B} \tag{4.30}$$

$$\frac{d^2\log f(X|\pi)}{d\pi^2} = -\frac{A(X)^2}{[(1-\beta)^X\beta + \pi A(X)]^2} + \frac{B^2}{[1-(1-\beta)^{k+1} - \pi B]^2} \tag{4.31}$$

The Fisher information is $I(\pi_0) = -\mathbb{E}\left[\frac{d^2\log f(X|\pi)}{d\pi^2}|_{\pi=\pi_0}\right]$, where π_0 is the true parameter. Obtaining an analytic form is difficult, but since X has finite state space and has probability mass function defined as in equation (4.7), we can compute this expectation exactly.

$$\begin{aligned} -\mathbb{E}\left[\frac{d^2\log f(X|\pi)}{d\pi^2}\Big|_{\pi=\pi_0}\right] &= -\sum_{x=0}^{k}\frac{d^2\log f(x|\pi)}{d\pi^2}\Big|_{\pi=\pi_0} f(x|\pi_0) \\ &= \sum_{x=0}^{k}\left\{\left[\frac{A(x)^2}{[(1-\beta)^x\beta + \pi A(x)]^2} - \frac{B^2}{[1-(1-\beta)^{k+1} - \pi B]^2}\right]\right. \\ &\quad \left.\cdot\frac{(1-\beta)^x\beta + \pi A(x)}{1-(1-\beta)^{k+1} - \pi B}\right\} \end{aligned} \tag{4.32}$$

Similarly, the standard deviation based on the CRLB of π^* is computed using Fisher Information, $\sigma = \sqrt{\frac{1}{nI(\pi_0)}}$.

Asymptotic Analysis We would like to obtain an asymptotic expression of equation (4.32) in terms of the truncation value k when $k\alpha, k\beta << 1$. Let us denote the first and second derivative of $f(x|\pi)$ respect to π as $f'(x|\pi)$ and $f''(x|\pi)$. The Fisher Information for the hybrid GRV can be written in the following form,

$$\begin{aligned} & -\mathbb{E}\left[\log''(f(X|\pi))|_{\pi=\pi_0}\right] \\ = & \sum_{x=0}^{k} \frac{f'(x|\pi)^2 - f''(x|\pi)f(x|\pi)}{f(x|\pi)^2} f(x|\pi)\Big|_{\pi=\pi_0} \\ = & \sum_{x=0}^{k} \left[\frac{f'(x|\pi)^2}{f(x|\pi)} - f''(x|\pi)\right]\Big|_{\pi=\pi_0} \end{aligned} \tag{4.33}$$

We would like to obtain an asymptotic expression of $f(x|\pi)$, $f'(x|\pi)$ and $f''(x|\pi)$. Let us start with $f(x|\pi)$. Assume $\beta = \omega\alpha$, α and β are in the same order of magnitude. The numerator of $f(x|\pi)$ can be written in the following,

$$\begin{aligned} & (1-\beta)^x\beta + \pi[(1-\alpha)^x\alpha - (1-\beta)^x\beta] \\ = & [\omega\alpha - x\omega^2\alpha^2 + \frac{1}{2}x(x-1)\omega^3\alpha^3](1-\pi) + [\alpha - x\alpha^2 + \frac{1}{2}x(x-1)\alpha^3]\pi + O(k^3\alpha^4) \\ = & [(1-\pi)\omega + \pi]\alpha - [\omega^2(1-\pi)+\pi]x\alpha^2 + \frac{1}{2}[\omega^3(1-\pi)+\pi]x(x-1)\alpha^3 + O(k^3\alpha^4) \\ = & t_1\alpha - t_2x\alpha^2 + \frac{1}{2}t_3x(x-1)\alpha^3 + O(k^3\alpha^4) \end{aligned} \tag{4.34}$$

where t_1, t_2, and t_3 are defined as the following,

$$\begin{aligned} t_1 &= \omega(1-\pi) + \pi \\ t_2 &= \omega^2(1-\pi) + \pi \\ t_3 &= \omega^3(1-\pi) + \pi \end{aligned} \tag{4.35}$$

Following the similar procedure, the denominator of $f(x|\pi)$ can be written as the following expression,

$$\begin{aligned} & 1 - (1-\beta)^{k+1} - \pi[(1-\alpha)^{k+1} - (1-\beta)^{k+1}] \\ = & t_1(k+1)\alpha - \frac{1}{2}t_2(k+1)k\alpha^2 + \frac{1}{6}t_3(k+1)k(k-1)\alpha^3 + O(k^4\alpha^4) \end{aligned} \tag{4.36}$$

Combining these two and using negative binomial expansion on the denominator, we have the following,

$$\begin{aligned} f(x|\pi) &= \frac{t_1\alpha - t_2 x\alpha^2 + \frac{1}{2}t_3 x(x-1)\alpha^3 + O(k^3\alpha^4)}{t_1(k+1)\alpha - \frac{1}{2}t_2(k+1)k\alpha^2 + \frac{1}{6}t_3(k+1)k(k-1)\alpha^3 + O(k^4\alpha^4)} \\ &= \frac{1}{(k+1)t_1}\frac{t_1 - t_2 x\alpha + \frac{1}{2}t_3 x(x-1)\alpha^2 + O(k^3\alpha^3)}{1 - \frac{1}{2}\frac{t_2}{t_1}k\alpha + \frac{1}{6}\frac{t_3}{t_1}k(k-1)\alpha^2 + O(k^3\alpha^3)} \\ &= \frac{1}{k+1}[1 - \frac{t_2}{t_1}x\alpha + \frac{1}{2}\frac{t_2}{t_1}k\alpha + \frac{1}{2}\frac{t_3}{t_1}x(x-1)\alpha^2 - \frac{1}{2}(\frac{t_2}{t_1})^2 xk\alpha^2 \\ &\quad -\frac{1}{6}\frac{t_3}{t_1}k(k-1)\alpha^2 + \frac{1}{4}(\frac{t_2}{t_1})^2 k^2\alpha^2 + O(k^3\alpha^3)] \end{aligned} \tag{4.37}$$

Note that t_1, t_2 and t_3 are function of π. Let us define the following,

$$\begin{aligned} g_1(\pi) &= \frac{t_2}{t_1} = \frac{\omega^2(1-\pi)+\pi}{\omega(1-\pi)+\pi} \\ g_2(\pi) &= \frac{t_3}{t_1} = \frac{\omega^3(1-\pi)+\pi}{\omega(1-\pi)+\pi} \end{aligned} \tag{4.38}$$

Now $f(x|\pi)$ becomes

$$\begin{aligned} f(x|\pi) &= \frac{1}{k+1}[1 - g_1(\pi)x\alpha + \frac{1}{2}g_1(\pi)k\alpha + \frac{1}{2}g_2(\pi)x(x-1)\alpha^2 - \frac{1}{2}g_1^2(\pi)xk\alpha^2 \\ &\quad -\frac{1}{6}g_2(\pi)k(k-1)\alpha^2 + \frac{1}{4}g_1^2(\pi)k^2\alpha^2 + O(k^3\alpha^3)] \end{aligned} \tag{4.39}$$

We can compute the asymptotic expression of $f'(x|\pi)$ and $f''(x|\pi)$ and substitute them into equation (4.33) to obtain the asymptotic express for the Fisher Information. Let us look at the first term $\sum_{x=0}^{k}\frac{f'(x|\pi)^2}{f(x|\pi)}$.

$$\begin{aligned} f'(x|\pi)^2 &= \frac{1}{(k+1)^2}[g_1'^2(\pi)x^2\alpha^2 + \frac{1}{4}g_1'^2(\pi)k^2\alpha^2 - g_1'^2(\pi)kx\alpha^2 + O(k^3\alpha^3)] \\ \frac{1}{f(x|\pi)} &= (k+1)(1+O(k\alpha)) \\ \frac{f'(x|\pi)^2}{f(x|\pi)} &= \frac{1}{k+1}[g_1'^2(\pi)x^2\alpha^2 + \frac{1}{4}g_1'^2(\pi)k^2\alpha^2 - g_1'^2(\pi)kx\alpha^2 + O(k^3\alpha^3)] \end{aligned}$$

$$\sum_{x=0}^{k}\frac{f'(x|\pi)^2}{f(x|\pi)} = \frac{1}{6}g_1'^2(\pi)k(2k+1)\alpha^2 - \frac{1}{4}g_1'^2(\pi)k^2\alpha^2 + O(k^3\alpha^3) \quad (4.40)$$

$$= \frac{1}{12}g_1'^2(\pi)k^2\alpha^2 + O(k\alpha^2) + O(k^3\alpha^3) \quad (4.41)$$

Next, let us look at the second term $\sum_{x=0}^{k} f''(x|\pi)$

$$\begin{aligned} f''(x|\pi) = \frac{1}{k+1}\{ & -g_1''(\pi)x\alpha + \frac{1}{2}g_1''(\pi)k\alpha + \frac{1}{2}g_2''(\pi)x(x-1)\alpha^2 \\ & -[g_1'^2(\pi) + g_1(\pi)g_1''(\pi)]xk\alpha^2 - \frac{1}{6}g_2''(\pi)k(k-1)\alpha^2 \\ & +\frac{1}{2}[g_1'^2(\pi) + g_1(\pi)g_1''(\pi)]k^2\alpha^2 + O(k^3\alpha^3)\} \end{aligned} \quad (4.42)$$

$$\begin{aligned} \sum_{x=0}^{k} f''(x|\pi) &= \frac{1}{2}g_2''(\pi)[\frac{k(2k+1)}{6} - \frac{k}{2}]\alpha^2 - \frac{1}{6}g_2''(\pi)k(k-1)\alpha^2 + O(k^3\alpha^3) \\ &= O(k\alpha^2) + O(k^3\alpha^3) \end{aligned} \quad (4.43)$$

Subtract the result of equation (4.43) from that of equation (4.41) and evaluate $g_1(\pi)$ at $\pi = \pi_0$, we have the expression for the approximated Fisher Information of the hybrid GRV with truncation k, as written in the following

$$\begin{aligned} & -\mathbb{E}\left[\log''(f(X|\pi))|_{\pi=\pi_0}\right] \\ = & \sum_{x=0}^{k}\left[\frac{f'(x|\pi)^2}{f(x|\pi)} - f''(x|\pi)\right]\Big|_{\pi=\pi_0} \\ = & \frac{1}{12}g_1'^2(\pi)|_{\pi=\pi_0}k^2\alpha^2 + O(k\alpha^2) + O(k^3\alpha^3) \\ = & \frac{1}{12}\frac{(\omega-1)^2\omega^2}{[\omega(1-\pi_0)+\pi_0]^4}k^2\alpha^2 + O(k\alpha^2) + O(k^3\alpha^3) \end{aligned} \quad (4.44)$$

Let us use the leading order term of the Fisher Information of the hybrid GRV with truncation to approximate the number of data points we need in order to obtain an accurate estimator of π_0. Given n data points and the true parameter π_0, we would

like to obtain an estimator with standard deviation $\epsilon\pi_0$. We can solve the following equation to get n.

$$\sqrt{\frac{12}{ng_1'^2(\pi)|_{\pi=\pi_0}k^2\alpha^2}} = \epsilon\pi_0$$
$$n = \frac{12}{g_1'^2(\pi)|_{\pi=\pi_0}\pi_0^2}\frac{1}{\epsilon^2k^2\alpha^2} \tag{4.45}$$

The first term $\frac{12}{g_1'^2(\pi)|_{\pi=\pi_0}\pi_0^2}$ is a function of ω and π_0. Let us denote it as $h(\omega, \pi_0)$. For $\omega = [0, 5]$ and $\pi_0 = [0, 1]$, $h(\omega, \pi_0)$ is plotted in $\log_{10}$ scale in Figure 4.2.

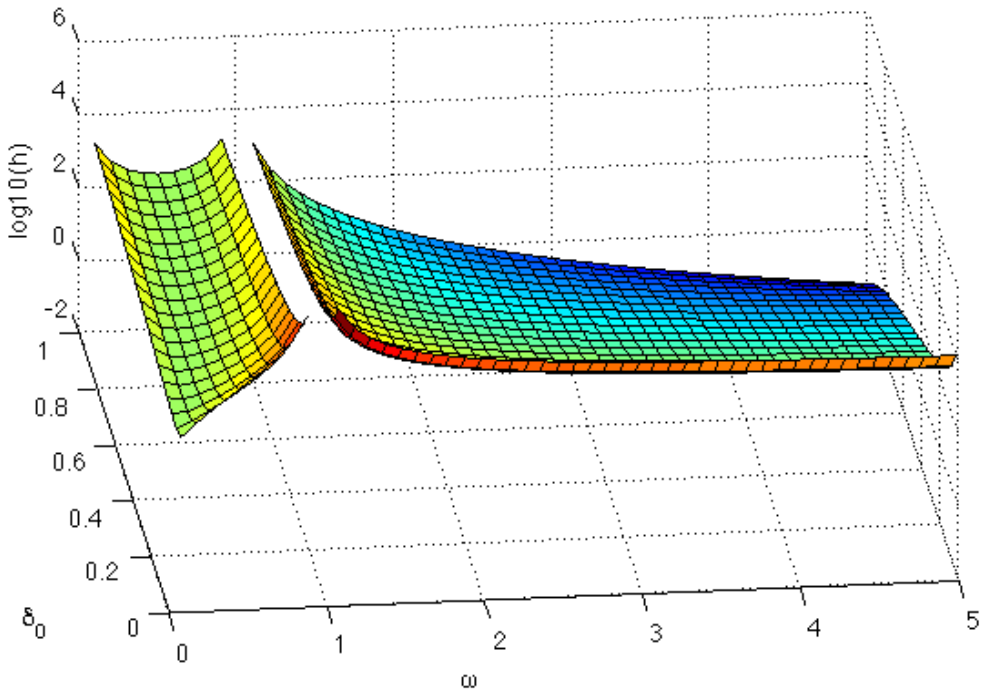


Figure 4.2: The values of the first term in equation (4.45): $h(\omega, \pi_0)$ (in $\log_{10}$ scale) in terms of different values of π_0 and ω.

The actual number of data point needed to achieve an estimator of standard deviation $\epsilon\pi_0$ can be computed by dividing $h(\omega, \pi_0)$ by $\epsilon^2k^2\alpha^2$. With the $k\alpha = 0.05$ and $\epsilon = 0.1$, n is plotted in $\log_{10}$ scale for the different values of ω and π_0 in Figure 4.3. Moreover, $\omega = 1$ implies $\alpha = \beta$, in which case the data set contains no information about the π_0. Therefore, $n = \infty$ when $\omega = 1$ as shown in the plot.

For example, when $\omega = 0.5$, $\pi_0 = 0.1$ (these values are chosen to be similar to the ones in the simulations later), we need approximately 7×10^9 data points in order to obtain an estimator of π with standard deviation of $0.1\pi_0$. The model is a simplified analogy for the case that an foreign-born individual entered with latent infection with probability π. If the individual has latent infection, he/she will have a probability α per month to be diagnosed. If he/she was susceptible at entry, then there is a probability β per month to be diagnosed. Based on the similar reason as in the toy model with single truncated GRV, we ignore the time between progression to active status and diagnosis. Here we are trying to estimate π from the data. This model illustrated that again we do not have enough data to obtain an accurate estimation.

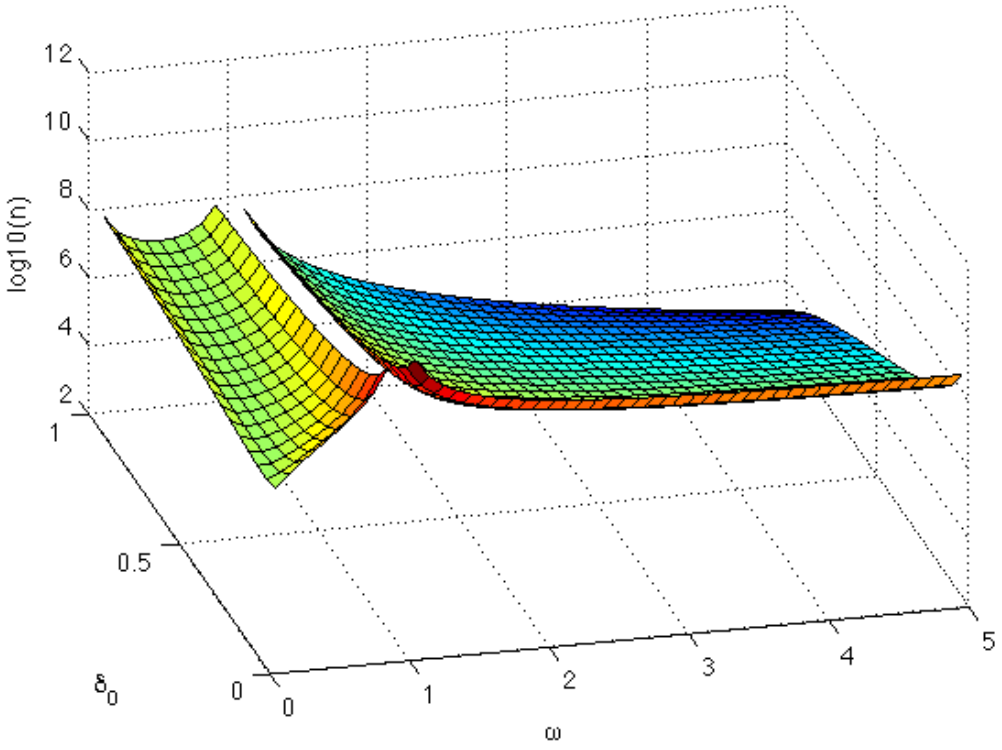


Figure 4.3: Given $k\alpha = 0.05$, $\epsilon = 0.1$, the number of data points needed n (in $\log_{10}$ scale) to obtain an estimator for π with standard deviation $0.1\pi_0$ in terms of different values of π_0 and ω. Note that $n = \infty$ when $\omega = 1$.

4.5 Numerical Examples

In order to demonstrate the effects of the truncation on the parameters estimation, numerical examples are performed.

4.5.1 Single GRV with Truncation

Let X be the GRV with success probability α and truncation value k and $\{x_i\}_{i=1}^n$ be the n observations. Again, X models the length of time that a foreign-born individual spends from entry to diagnosis. k (months) is the life span of an individual, starting from the time of entry. Therefore, individuals exist for more than k months in our model will be removed. This effect is modeled by the truncation at k. The log-likelihood function, $\log\mathcal{L}(\alpha)$, is defined as in equation (4.4). $\log\mathcal{L}(\alpha)$ is computed for a range of different values of $\alpha \in \boldsymbol{\varphi}$. The α which maximizes $\log\mathcal{L}(\alpha)$ is chosen as the maximum likelihood estimator.

$$\alpha^* = \max_{\alpha\in\boldsymbol{\varphi}} \log\mathcal{L}(\alpha) \tag{4.46}$$

This experiment tries to mimic the scenario of estimating the probability of becoming active per month given a patient is latently infected, which is represented by α in this experiment. Since the probability of being diagnosed once a patient become active $\gamma >> \alpha$, let us assume the time that a patient spends from entry to diagnosis is modeled by a GRV with success probability α and truncated at k. 3 different values for k are chosen to demonstrate the impact of the truncation on the estimation of α. First, a large k is chosen, which has a cdf, $F_Y(k)$, greater than 99%. In this case, the mle accurately reveals the real parameter. Second, a relatively smaller k is chosen, with cdf around 45%. In this case, the mle estimator is still accurate, but the standard error increases. Lastly, a small k with cdf less than 5% is chosen. In this case, the mle fails to recover the real parameter and the standard deviation is large (about the same as α, the parameter itself).

The estimation is performed under 3 different parameter settings. In all of these 3 settings: $\alpha_0 = 1\times10^{-4}$, $\boldsymbol{\varphi}$ is a vector whose entries are 100 evenly separated points in $[1\times10^{-9}, 3\times10^{-4}]$. Under each setting, 5000 random variables are sampled for 1000 times and α^* is computed each time. The histogram of the 1000 α^* is shown

in Figure 4.4

- Setting 1:
  $k = 50000$, $F_Y(k) = 0.9933$
  mean of α^*: 1.00×10^{-4}, $\sigma(\alpha^*) = 1.84 \times 10^{-6}$
  $\sigma = 1.55 \times 10^{-6}$

  5000 GRVs with truncation $k = 50,000$ are generated. Recall that Y is the GRV without the truncation. The cdf of Y at k $F_Y(k) = 0.9933$ implies that the truncation chooses 99.33% of Y. The standard deviation of $\alpha^* = 1.84 \times 10^{-6}$. The standard deviation computed based on the CRLB $\sigma = 1.55 \times 10^{-6}$.

- Setting 2:
  $k = 6000$, $F_Y(k) = 0.4513$
  mean of α^*: 9.97×10^{-5}, $\sigma(\alpha^*) = 8.71 \times 10^{-6}$
  $\sigma = 8.23 \times 10^{-6}$

- Setting 3:
  $k = 520$, $F_Y(k) = 0.0508$
  mean of α^*: 1.07×10^{-4}, $\sigma(\alpha^*) = 8.21 \times 10^{-5}$
  $\sigma = 9.40 \times 10^{-5}$

Note that in the setting 3, standard deviation computed based on CRLB is greater than the actual standard deviation of the 1000 estimated α^*. This is because in the computation, α is chosen from $[1 \times 10^{-9}, 3 \times 10^{-4}]$. A numerical example is performed to demonstrate the effect of this clipping. 1000 random variables following the normal distribution of $\mathcal{N}(1 \times 10^{-4}, (9.40 \times 10^{-5})^2)$ are generated: $\{z_i\}_{i=1}^{1000}$. Set $z_i = 1 \times 10^{-9}$ if $z_i < 1 \times 10^{-9}$ and $z_i = 3 \times 10^{-4}$ if $z_i > 3 \times 10^{-4}$. The mean of the resulting $\{z_i\}_{i=1}^{1000}$ is 1.09^{-4} and standard deviation is 8.06×10^{-5}. The histogram of the resulting $\{z_i\}_{i=1}^{1000}$ are shown in Figure 4.5.

We also computed the standard deviation based on the CRLB with 5000 observations. With the success probability as 1×10^{-4}, σ for different values of k is plotted in Figure 4.6. As shown in the plot, σ increases as the truncation value k

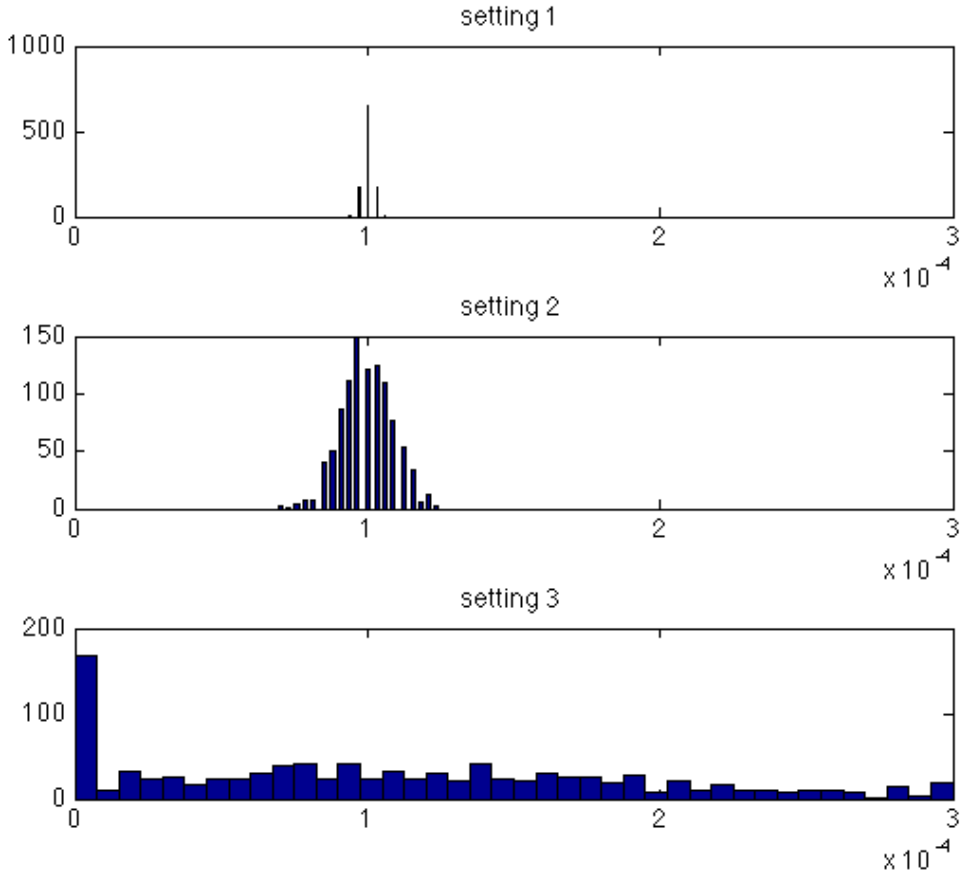


Figure 4.4: The histograms of α^* in 1000 estimations with 3 different settings. The true value α_0 is 10^{-4}

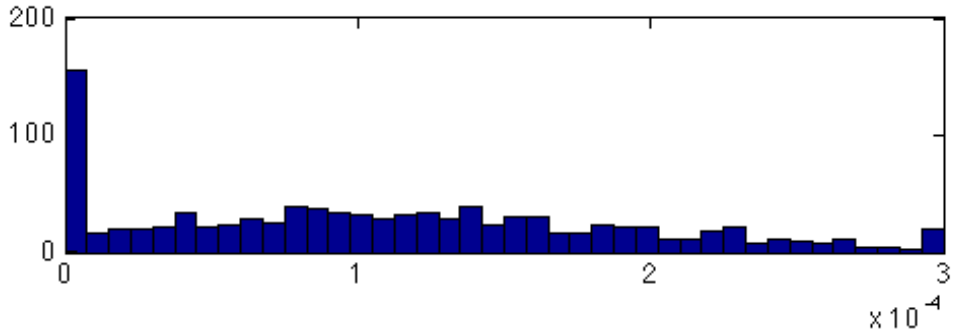


Figure 4.5: The histogram of the 1000 random variables sampled from normal distribution $\mathcal{N}(1 \times 10^{-4}, (9.40 \times 10^{-5})^2)$ and values clipped between $[1 \times 10^{-9}, 3 \times 10^{-4}]$. The mean is 1.09×10^{-4} and sd is 8.06×10^{-5}

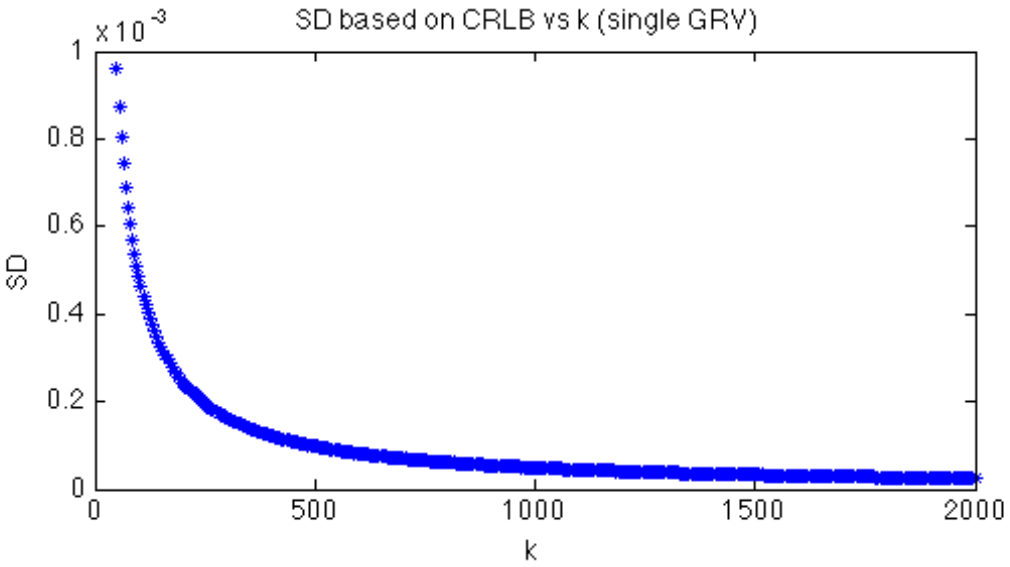


Figure 4.6: Standard deviation of α^* based on the CRLB with 5000 observations is plotted against different truncation values k. Success probability: 1×10^{-4}

decreases. The data set we have (under the assumptions of our model) is most similar to case 3 with α corresponds to the probability of becoming active from latent. The experiment showed that with 5000 data points, which is more than the data points we have, the standard error we have for the estimator of α is greater than α itself. This indicates that it is unfeasible to obtain an accurate estimation for α using our data set.

4.5.2 Hybrid GRV with Truncation

The estimation of the π in the hybrid GRV follows the same procedure as the single GRV with truncation. The log-likelihood function $\log\mathcal{L}(\pi)$ is defined as in equation (4.9). $\pi^* = \max_{\pi\in\varphi} \log\mathcal{L}(\pi)$.

This experiment investigates the scenario of estimating the probability of an immigrant entering the country with latent infection. Given a foreign born person at entry, there is a probability π he/she is latently infected. In this case he/she will have a probability α of becoming active each month. On the other hand, the patient has a probability $1-\pi$ of not being latently infected. In this case, he/she has a probability β of being infected each month. Here we simplify the transmission

by assuming patient becomes active immediately after infection. Similar as before, we assume the time from becoming active to diagnosis is negligible. Therefore, for each patient, the time from entry to diagnosis is modeled by this hybrid GRV with truncation k. Using the same procedure as before, three values of k are chosen to explore the effect of truncation on the estimation.

5000 hybrid GRV with truncation are simulated 1000 times. The estimation is perform in each simulation with each of the 3 parameters settings. In all of the 3 settings, $\alpha = 1 \times 10^{-4}$,$\beta = 5 \times 10^{-5}$, $\pi_0 = 0.1$, $\boldsymbol{\varphi}$ is a vector whose entries are 100 evenly separated points in $[0, 1]$. The histograms of the estimated π are shown in Figure 4.7.

- Setting 1:
  $k = 100000$, $F_Y(k) = 0.9939$
  mean of $\pi^* = 0.1004$, $\sigma(\pi^*) = 0.0252$
  $\sigma = 0.0246$.

  5000 hybrid GRVs with truncation value $k = 50000$ are generated. Note that Y is the hybrid GRV without the truncation. $F_Y(k) = 0.9939$ implies the truncation chooses 99.39% of the orignally generated Y. The mean of the 1000 estimations of π is 0.1004, the standard deviation is 0.0252. The standard deviation computed based on the CRLB is 0.0246.

- Setting 2:
  $k = 11000$, $F_Y(k) = 0.4474$
  mean of $\pi^* = 0.1045$, $\sigma(\pi^*) = 0.0630$
  $\sigma = 0.0647$.

- Setting 3:
  $k = 930$, $F_Y(k) = 0.0498$
  mean of $\pi^* = 0.3560$, $\sigma(\pi^*) = 0.4173$
  $\sigma = 0.6462$.

Under the assumption of our model, the data set we have is similar to the case 3. π corresponds to the probability of a foreign-born individual being latent

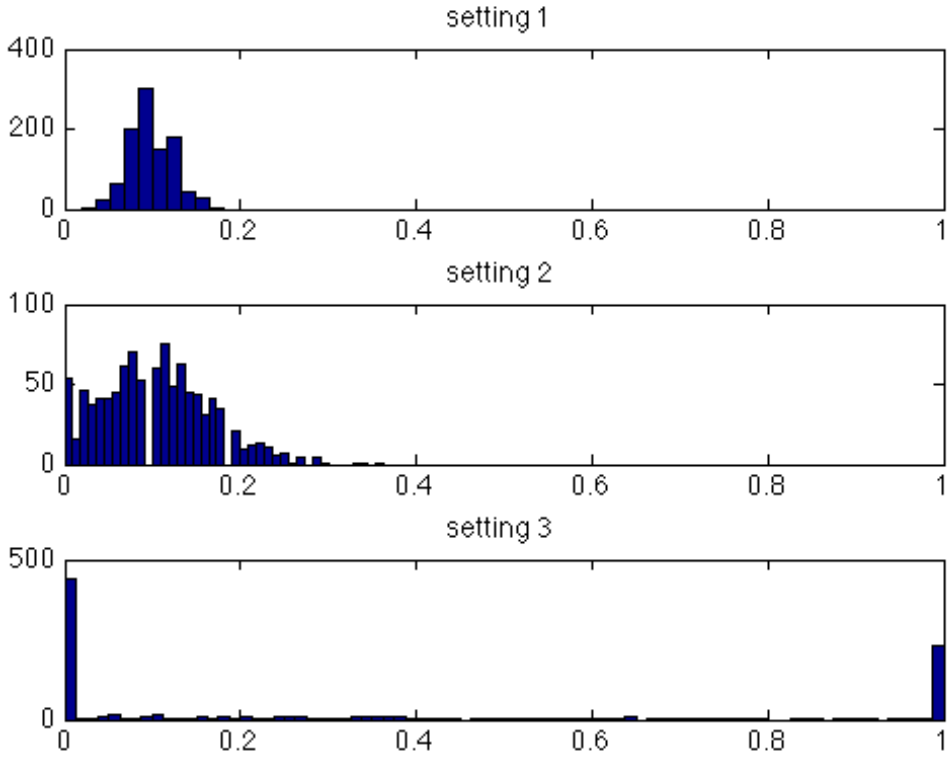


Figure 4.7: The histograms of π^* in 1,000 estimations with 3 different settings. The true value: π_0 is 0.1.

at entry. For an individual, the probability of being diagnosed is α per month if latent and β per month if susceptible. The percentage of individuals with latent infections to develop active TB is 5-10% in their life time [6]. Here we assume the probability for an individual to be diagnosed with TB before reaching the life span k is 5%. Here we are trying to estimate π. The experiments shows that, with 5000 data points, the standard error of the estimator of π is again larger than its value. Therefore, getting an accurate estimation based on our data set is impossible.

4.6 Conclusion

In this section we investigated the feasibility of parameter estimation with the real patient data. The data for foreign born patients, containing entry and diagnosis times, are collected under the condition that they are diagnosed before

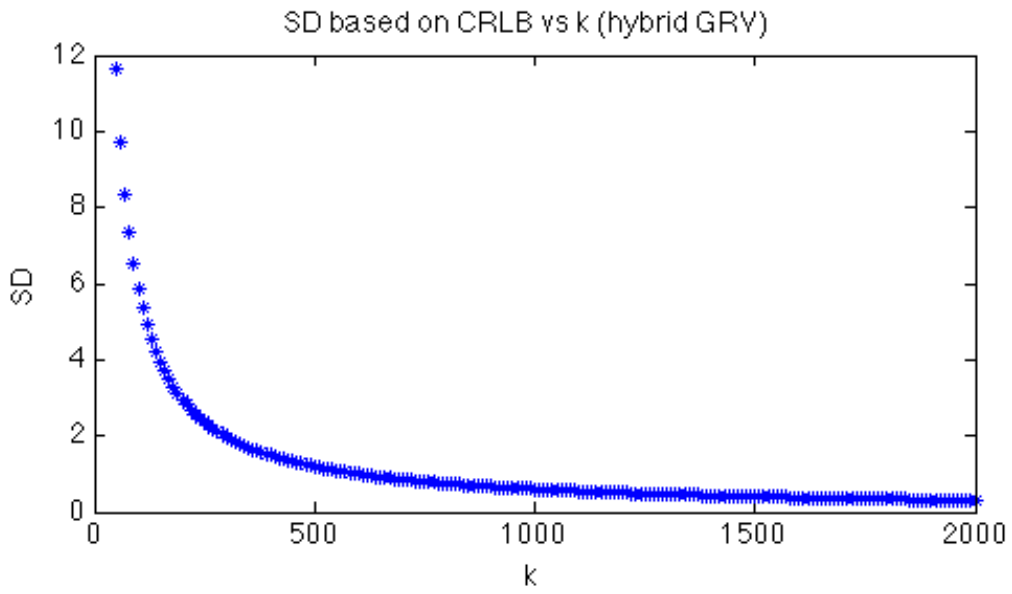


Figure 4.8: Standard deviation of π^* based on the CRLB with 5000 observations plotted against different truncation values k. With $\alpha = 1 \times 10^{-4}$, $\beta = 5 \times 10^{-5}$, $\pi = 0.1$.

they are removed from the population by death (due to natural causes). The time from entry to diagnosis is modeled as random variables. The effect of the limited life span truncates the random variables at k. We simplified the transmission dynamics and model the time from entry to diagnosis by geometric/hybrid geometric random variables with truncation value k. We chose three different values of k to simulate data and performed maximum likelihood estimations on the parameters. It is found that the standard deviation of the maximum likelihood estimator increases as k decreases. When k is set at a small values such that around 5% of the random variables is less than k, the standard deviation of the estimators are approximately the same as the estimators themselves. Therefore the estimators in the scenarios with small truncation values are highly unreliable. Note that random variables with small truncation values is a good approximation of the time that real TB patients spend from entry to diagnosis, since only around 5-10% of patients with latent infections become active in their lifetime [6]. In the toy model with single truncated GRV, the success probability α represents the probability that an latent individual becoming active in a month. In the model with hybrid GRV, β represents the probability for a susceptible individual to be diagnosed each month (α has the

same meaning as in the single GRV model). π represents the probability of entering with latent infection.

The characteristics of the data posed difficulties on parameter estimation. Fortunately, our model is shown to be insensitive to parameters. In the following chapter, a range of parameters are chosen and our model is tested with different combinations of these parameters. Finally, the model is applied to the real data to help identify recent transmission TB cases.

CHAPTER 5
Application of the TB spread Model

5.1 Introduction

In order to better control TB and identify outbreaks, we need to distinguish endogenous reactivations from recent transmissions. The traditional epidemiological approach does so by interviewing patients to identify transmission routes. This approach is both labor intensive and time consuming. The TB genotyping technologies have enriched the traditional methods. By clustering TB patients into smaller groups, it makes probabilistic modeling of the transmission routes possible. Our model takes advantage of the smaller sizes of the patient clusters and provides a probabilistic estimation of endogenous reactivation versus recent transmission. We first tested our model in the simulation data. The sensitivity to parameters is also evaluated by experimenting with different combination of parameters values. Finally, the model is applied to the New York City data collected by Center of Disease Control (CDC) from 2000-2007.

We use the receiver operating characteristic (ROC) curve to evaluate the models' performance. Before proceeding to the application of our model, we will introduce the ROC curve in the following section.

5.2 Receiver Operating Characteristic Curve

A receiver operating characteristic (ROC) curve is a plot that evaluates the performance of a binary classifier. It is firstly used in signal detection theory and is now a common technique in statistics and machine learning. It plots the ratio of *true positive* out of the total positive against the ratio of *false positive* out of the total negative at different discrimination thresholds [48].

5.2.1 Binary Classification Problem

Let's consider a binary classification problem. There are N instances, each belonging to one of two classes: *positive* (P) or *negative* (N). A classification model

is a mapping of the instances to predicated classes [48]. The classification models will produce a score, which is used to determinate the classes of the instances. Some of models' output scores are discrete and others might be continuous. For models with continuous score, a threshold is chosen to discriminate different classes.

An example of binary classification with continuous score will be the pregnancy test. The urine concentration of human chorionic gonadotropin (hCG), a hormone produced by the fertilized egg, can be used to determine whether the test subject is pregnant or not [49]. A hCG concentration higher than certain level will be classified as pregnant. Each test subject will be an instance, belonging to one of two classes: pregnant(P) or not pregnant(N). The hCG concentration is the score. A certain level of the hCG concentration (the threshold) needs to be chosen to assign a class to each subject.

Given a classification model and an instance, there are four possible outcomes.

1. The true class is P, and is classified as P. This case is called *True Positive*
2. The true class is N, and is classified as P. This case is called *False Positive*
3. The true class is N, and is classified as N. This case is called *True Negative*
4. The true class is P, and is classified as N. This case is called *False Negative*

These four outcomes are usually presented in a confusion matrix, shown in table (5.1)

Table 5.1: A confusion matrix containing the four possible outcomes of a binary classification problem

		True Class: P	True Class: N
Classified Class	P	True Positive	False Positive
	N	False Negative	True Negative
		Total Positive	Total Negative

5.2.2 ROC Curve

There are two terms associated with ROC analysis: *True Positive Rate* (TPR) and *False Positive Rate* (FPR). They are computed as following

$$\text{TPR} = \frac{\text{Number of True Positive}}{\text{Number of Total Positive}} \tag{5.1}$$

$$\text{FPR} = \frac{\text{Number of False Positive}}{\text{Number of Total Negative}} \tag{5.2}$$

For different values of threshold, we will have different values of TPR and FPR. A classification model with a specific threshold value which correctly identifies 60% of the positive instances but also incorrectly classifies 30% of the negative instances as positive will result in a point at $[0.6, 0.3]$ on the ROC curve. A model of random guess will generate a pair of TPR and FPR which lands on the straight line of $y = x$. This is because at any threshold, if the random guess has a probability p to classify a positive instance as positive, it will have the same probability to identify a negative instance as positive.

Suppose there are N instances in a problem and each instance is assigned a score s_i by the binary classification model. We sort the instances to make the scores in descending order, i.e. $s_1 \geq s_2 \geq ... \geq s_N$. The threshold is set to equal to each s_i. Instances with a score greater than the threshold will be classified as positive, otherwise as negative. The corresponding TPR(s_i) and FPR(s_i) are computed according to equation (5.1) and (5.2). The ROC curve is generated by plotting TPR(s_i) versus FPR(s_i). Since there are no instances with score greater than s_1 and all the instances will have score greater than s_N, we have the following: TPR($s_N - \epsilon$) = FPR($s_N - \epsilon$) = 1 and TPR(s_1) = FPR(s_1) = 0 (ϵ is a dummy positive value to create a point at (1,1)). Note TPR and FPR are both ratios, therefore both of the domain and the range of the ROC curve will be $[0, 1]$.

The ROC curve visualizes the trade-off between the benefit (TPR) and the cost (FPR) of a classification model. A good model will generate a ROC curve above the $y = x$ line. The following example shows an classification problem with 10 positive and 10 negative instances and the corresponding ROC curve. The positive

instances are represented as "1"s while the negatives ones are represented as "0"s. The classification model generates a score from 0 to 1. This result is shown in table (5.2). Figure (5.1) shows the ROC curve for this classification problem.

Table 5.2: Example of a classification problem with 10 positive (represented as 1) and 10 negative (represented as 0) instances. The classification model generates a score from 0 to 1. The instances are sorted in descending order by their assigned score and numbered 1 to 20.

ID	True Class	Model Score	ID	True Class	Model Score
1	1	0.97	11	1	0.50
2	1	0.94	12	0	0.47
3	1	0.92	13	1	0.43
4	1	0.86	14	0	0.35
5	1	0.83	15	1	0.32
6	0	0.83	16	0	0.23
7	1	0.77	17	0	0.21
8	1	0.63	18	0	0.14
9	0	0.56	19	0	0.08
10	0	0.55	20	0	0.03

5.2.3 Area Under The Curve (AUC)

It will be useful to condense the information of a ROC plot into a single scalar value, a statistic. In this way, we could compare different classification models. One common statistic used here is the area under the ROC curve (AUC). Since the limits of both variables of the ROC plot are $[0, 1]$ and AUC is a portion of an unit square, AUC takes values from 0 to 1. The larger the AUC, the better average performance of the classification model. AUC is the probability that the model will give a randomly chosen positive instance a higher score than a randomly chosen negative one [48].

Given a model, assume we already computed its ROC curve, we will have the following:

- scores it generates for each instances $\{s_i\}_{i=1,2,...,N}$ (we sorted the instances to make $s_1 \geq s_2 \geq ... \geq s_N$)

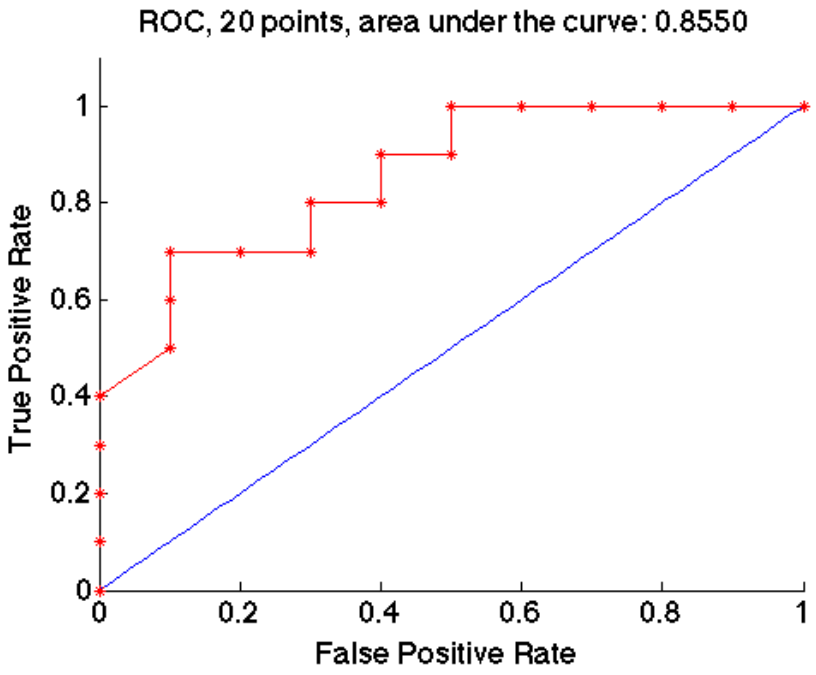


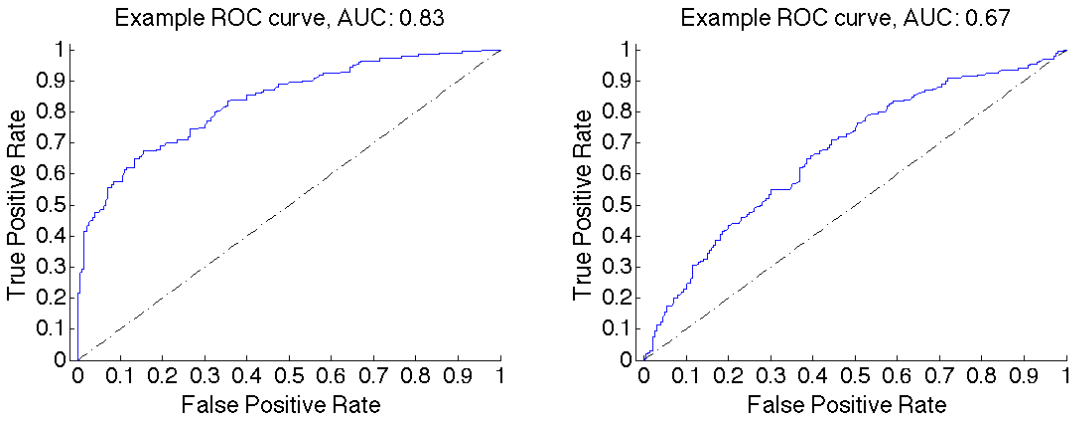
Figure 5.1: ROC curve for the example in section (5.2.2). There 10 positive and 10 negative instances, each is assigned a score by the classification model. From lower left to upper right, each point represents the TPR and FPR values computed with thresholds s_1, s_2, ..., s_{20}.

- The TPR(s_i) and FPR(s_i) are computed at each threshold (chosen to be s_i)
- TPR(s_i) is plotted against FPR(s_i) to generate the ROC curve.

The AUC can be computed by integrating TPR(s_i) against FPR(s_i). We will use the *Trapezoidal rule* here.

$$\text{AUC} = \sum_{i=1}^{N-1} \frac{1}{2} \left[\text{TPR}(s_i) + \text{TPR}(s_{i+1})\right] \left[\text{FPR}(s_i) - \text{FPR}(s_{i+1})\right] \tag{5.3}$$

Figure (5.2) shows an example of two binary classification models working on the same set of instances. The AUC of model A is 0.83 comparing to 0.67 for model B. This implies that model A has a better average performance than B.



(a) Model A, AUC: 0.83 **(b) Model B, AUC: 0.67**

Figure 5.2: There are two binary classification models, model A and B. They both work on the same instances. Model A(5.2a)yields a AUC of 0.83, while model B(5.2b) yields 0.67. This example indicates that model A has a better average performance than B.

5.3 Simulation

We first investigate the performance of our models on simulated data, using both the 2-body method and the 1-body mean field. Patients' entry/diagnosis times and the initial status (latent or susceptible) are simulated. The conditional probabilities of being susceptible at the time of entry of one patient given the entry and diagnosis times of every patient in the cluster are computed using our models. This conditional probability will be used as a score to determine whether this patient is infected after entering the country. Both the simulation and the computation will use the same set of parameters for consistency. The procedures of the simulation are listed as the following,

1. The size of foreign-born population is N_F, among whom there are three types of people, in terms of TB disease: Susceptible(S), Latent(L) and Active(A).

2. Foreign-born people enter their group through immigration. There is a certain immigration rate r_i people/month. There are two possible states for the immigrants when they enter: they have a probability $1-\pi$ to be susceptible and probability π to be latent.

3. Suppose N_F is the total number of foreign-born people in a given month and N_F^a active TB patients among them. The infectivity contributed by an active TB foreign-born patient is Γ, indicating that, for a susceptible person, the probability of being infected by the foreign-born population in a given month is $\frac{N_F^a}{N_F}\Gamma$. We assume the probability of being infected by the domestic bath is a constant, β_d. Therefore, for a susceptible foreign-born individual, the probability that he/she will be infected is $1-(1-\frac{N_F^a}{N_F}\Gamma)(1-\beta_d)$ in that given month.

4. Once an individual is infected, he/she has a probability δ of becoming active immediately or $1-\delta$ of acquiring latent infection. Patients with latent infection have probability α to become active per month. After an active patient is diagnosed, he/she will be removed from the population.

5. There is a life span for all the individuals: k. Anyone existing for a period longer than k will be removed from the population. This corresponds to death due to natural causes.

6. The simulation is run for time U in order to let the system reach a stationary state. The first c patients who are diagnosed after U are recorded as one cluster.

The algorithm for simulation is shown in Algorithm 1. The definitions of variables used in the algorithm are shown in the following list.

- I: a person with entry time, diagnosis time and entry status. The diagnosis time is set to -1, if the individual is not diagnosed yet.

- F_S, F_L, F_A, F_D: sets of foreign-born individuals who are susceptible, latent, active and diagnosed.

A summary of input and output in Algorithm 1, are shown in the Table 5.3.

5.3.1 Data and Parameter Selection

We have data of foreign-born TB patients in New York city who were diagnosed in the period of Nov 2000 - Dec 2007 [46]. The data is provided by Center of Disease

Table 5.3: A summary of the input and out variables in the Algorithm 1.

Input	Meaning
α	Probability of becoming active given latent (per month)
Γ	Infectivity contributed by a single active patient
β_d	Probability of being infected by domestic bath (per month)
δ	Probability of becoming active immediately after infection
γ	Probability of being diagnosed given active (per month)
c	Size of the cluster
Output	Meaning
F_D	A cluster of patients who are diagnosed

Control (CDC). For each patient in the data set, we have the entry and diagnosis times. We also have the Spoligotypes and RFLP of the patient's MTBC isolate. Patients are clustered based on Spoligotypes and RFLP. (Patients in one cluster share the identical Spoligotypes and RFLP.)

The original data have 5,258 patient records, including US-born patients who do not have entry time. Among the data, 2 are missing the Spoligotype values, 3 of them have diagnosis time prior to entry time and 1,512 of them omit the entry time information. Excluding those patients with missing information leaves us 3,741 patient records. The earliest diagnosis time is 29-Nov-2000, the latest is 28-Dec-2007. After clustering the patients according to their Spoligotype and RFLP, there are 2,679 clusters, including clusters with size 1. The distribution of the sizes of the clusters are shown in Figure 5.3. The clusters with size greater than 15 accounts for only 0.55% of the total clusters and they are not shown. Here is some basic

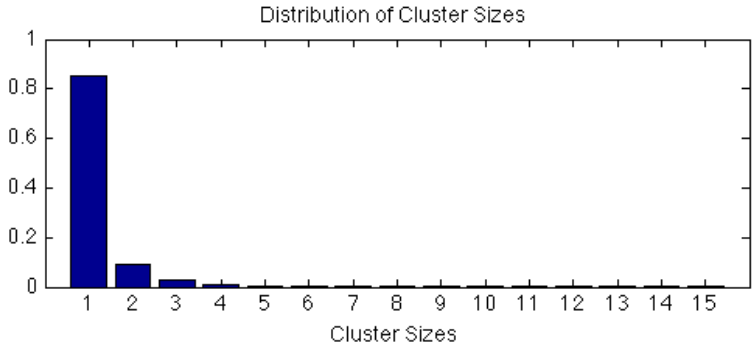


Figure 5.3: The distribution of the cluster sizes (smaller or equal to 15) of the NYC patient data.

Algorithm 1 Function SimAll

Function SimAll simulates the foreign-born patient clusters
Input: α, Γ, β_d, δ, γ, c
Output: F_D

function SimAll(T, α, Γ, β_d, δ, γ)
    set F_S, F_L, F_A to initial sizes
    set F_D to size 0
    for currentMonth = 1:T **do**
        add r_i number of I to F_S or F_L
            add I to F_L with probability π, I.status = L
            add I to F_S with probability $1-\pi$, I.status = S
        for all I, I.entryTime = currentMonth, I.diagnosisTime = -1
        for all I, if currentMonth $-$ I.entryTime $> k$, remove I.

        $\beta_f = 1 - \left(1 - \frac{\text{size}(F_A)}{\text{size}(F_S+F_A+F_L)}\Gamma\right)(1-\beta_d)$
        for each I in F_S **do**
            if rand()$\leq \beta_f$ AND rand()$\leq \delta$ **then**
                move I from F_S to F_A
            else if rand()$\leq \beta_f$ AND rand()$> \delta$ **then**
                move I from F_S to F_L
        for each I in F_L **do**
            if rand() $\leq \alpha$ **then**
                move I from F_L to F_A
        for each I in F_A **do**
            if rand() $\leq \gamma$ **then**
                move I from F_A to F_D
                I.diagnosisTime = currentMonth

    Return last c entries in F_D

information about clusters: 1) 84.96% have size 1; 2) 98.25% have size less than 5; 3) 99.29 % have size less than 10; 4) The largest cluster size is 44.

According to a census report on New York City [50], the city's population was 8,244,910 in 2011, among those 37.2% are foreign-born. 323,082 foreign-born people entered the city from 2000-2010. Simulation at this scale while at the same time studying the transmission dynamics simultaneously at the individual level are computationally expensive. Instead of simulating the real population, a model with

smaller population size is simulated. Our goal for the simulation is to produce approximately the same number of patients with a specific strain of TB with the NYC data, which is around 2 in 7 years. We assume the following: 1) the initial sizes of the susceptible, latent and active population for the foreign-born population are set to be 1000, 20, 1; 2) every month, 3 foreign-born people enter; 3) the life span is 600 months.

Once a patient develops active TB disease, he/she will develop symptoms like nausea, vomiting and fever [6]. We assume the average time for an active TB patients getting diagnosed is 3 months. Therefore, we have $\gamma = 0.3333$, which is the probability that an active patient is diagnosed in a given month. Also according to Centers for Disease Control and Prevention, "Overall, without treatment, about 5 to 10% of infected persons will develop TB disease at some time in their lives. About half of those people who develop TB will do so within the first two years of infection." [5]. We assume the chance of becoming active immediately after infection is 5%, i.e. $\delta = 0.05$.

At this stage we are uncertain about α, the probability per month to become active given the patient has latent infection; Γ, the infectivity per month contributed by a patient with active TB inside the cluster; π, the probability that an immigrant enters the country with latent infection of a specific strain of TB. Estimating these parameters from the data is impractical, as discussed in Chapter 4. We would like to choose parameters to sweep low-median-high possible values for these three parameters and test the sensitivity of our models to those values.

Several research works have been done on TB modeling of the foreign-born patients [51–55], in which the estimated values of α ranging from 2.63×10^{-5} to 8.16×10^{-5}. These estimations make no assumptions of the life span of foreign-born person. Since we assume a shorter life span for foreign-born persons (600 months), we choose slightly greater values for α: $5 \times 10^{-5}, 1 \times 10^{-4}, 2 \times 10^{-4}$. An active TB patient, on average, will cause 7 infections in a year [39]. Translating this into infectivity contributed by an active TB patient, we will have $\Gamma = 0.583$. Among all the TB cases in foreign-born patients, the percentage of recent transmission ranges from 14-24%, as indicated in [51, 53, 54]. In order to have the simulation generate

a similar percentage of recent transmission, we choose values of Γ as $0.5, 1, 2$. With the goal of generating around 2 cases in 7 years, values of π are chosen to be the following, $\pi : 0.035, 0.07, 0.2$. In summary, we have $3 \times 3 \times 3 = 27$ experiments to test (one for each set of parameters).

As discussed in Chapter 1, the TB incidence rate among foreign-born persons is much higher than the domestic population in United States. At this stage, we assume there is no infectivity from the domestic population, i.e. $\beta_d = 0$. It is found that the system becomes stationary after 1200 months for all the 27 experiments. Therefore, we choose $T = 1200$ and the first 5 TB cases after 1200 months are collected. The disease status at entry and entry/diagnosis times of these 5 patients are recorded. The simulations are run 1000 times for each experiment. Among all the TB cases, the percentage of diagnosed people entering the country susceptible and the average active patients per month are shown in Table 5.4.

Model Evaluation: In order to evaluate the performance of our model, we propose the following naïve method, which serves as a control of our models. For a susceptible foreign-born person, the naïve method assumes constant infectivity regardless of the existence of other active patients in the cluster. Let N_F^a be the average active patients per month and N_F be the average size of the foreign-born population; both are averaged over all the 1000 simulations. The probability of being infected is defined as following,

$$\beta_v = 1 - (1 - \frac{N_F^a}{N_F}\Gamma)(1 - \beta_d) \tag{5.4}$$

Note that β_v is a constant for the naïve method. For a patient who entered at t_0 with latent infection, the probability of he/she being diagnosed at t_1 is simply computed as

$$p(t_1|\overline{\mathcal{S}}, t_0) = h(t_1 - t_0, \alpha, \gamma) \tag{5.5}$$

where $h(x, \theta_1, \theta_2)$ is defined as in equation (3.5), it computes the probability mass function of the sum of two geometric random variables with success probability

Table 5.4: 1000 simulations are run for each of the 27 parameter settings. For each simulation, the first 5 cases after 1200 months are collected as one cluster. The first three columns are the values for the parameters. The 4th column, "% of $\mathcal{S}$" , is the percentage of diagnosed persons who entered the country susceptible and were infected domestically. The last column, "avg. act.", is the number of average active TB patients per month.

exp no.	α	β_c	π	% of $\mathcal{S}$	avg. act.
1	5×10^{-5}	0.5	0.035	0.0638	0.0139
2	1×10^{-4}	0.5	0.035	0.0652	0.0229
3	2×10^{-4}	0.5	0.035	0.0864	0.0440
4	5×10^{-5}	1.0	0.035	0.1236	0.0148
5	1×10^{-4}	1.0	0.035	0.1178	0.0247
6	2×10^{-4}	1.0	0.035	0.1660	0.0482
7	5×10^{-5}	2.0	0.035	0.2242	0.0167
8	1×10^{-4}	2.0	0.035	0.2468	0.0291
9	2×10^{-4}	2.0	0.035	0.3466	0.0666
10	5×10^{-5}	0.5	0.070	0.0568	0.0229
11	1×10^{-4}	0.5	0.070	0.0580	0.0442
12	2×10^{-4}	0.5	0.070	0.0816	0.0874
13	5×10^{-5}	1.0	0.070	0.1050	0.0241
14	1×10^{-4}	1.0	0.070	0.1250	0.0478
15	2×10^{-4}	1.0	0.070	0.1724	0.0977
16	5×10^{-5}	2.0	0.070	0.1970	0.0273
17	1×10^{-4}	2.0	0.070	0.2410	0.0576
18	2×10^{-4}	2.0	0.070	0.3434	0.1280
19	5×10^{-5}	0.5	0.200	0.0416	0.0625
20	1×10^{-4}	0.5	0.200	0.0622	0.1243
21	2×10^{-4}	0.5	0.200	0.0802	0.2447
22	5×10^{-5}	1.0	0.200	0.0866	0.0654
23	1×10^{-4}	1.0	0.200	0.1128	0.1321
24	2×10^{-4}	1.0	0.200	0.1484	0.2716
25	5×10^{-5}	2.0	0.200	0.1726	0.0743
26	1×10^{-4}	2.0	0.200	0.2124	0.1527
27	2×10^{-4}	2.0	0.200	0.2938	0.3319

θ_1 and θ_2. For patients who entered susceptible, the probability of he/she being diagnosed at t_1 is computed as follows,

$$p(t_1|\mathcal{S}, t_0) = \sum_{i=t_0}^{t_1} (1-\beta_v)^{i-t_0} \beta_v [\delta(1-\gamma)^{t_1-i}\gamma + (1-\delta)h(t_1-i, \alpha, \gamma)] \tag{5.6}$$

$$p(\mathcal{S}|t_0, t_1) = \frac{p(t_1|\mathcal{S}, t_0)\pi}{p(t_1|\overline{\mathcal{S}}, t_0)\pi + p(t_1|\mathcal{S}, t_0)(1-\pi)} \tag{5.7}$$

The naïve method simply tries to mimic the 1-body mean field. The only difference is that the infectivity remains constant regardless other patients in the cluster. Next we will test our models from different aspects.

Different Order: Let us denote the patients as A,B,C,D,E according to the order of the diagnosis times. An illustration of a cluster of size 5 is shown in Figure 5.4. Patient E is set as the target patient: the conditional probability of E being

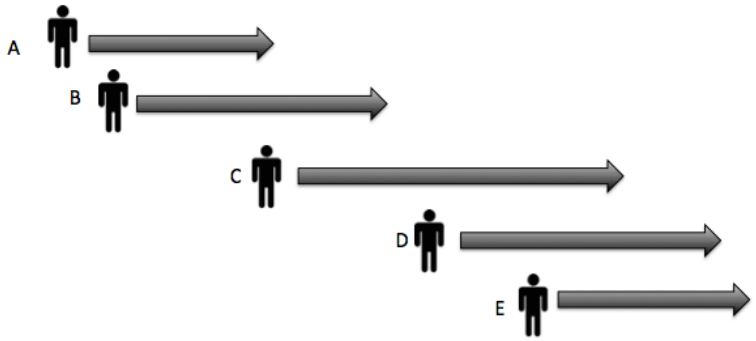


Figure 5.4: An illustration of a patient cluster of size 5. The patients are ordered according to the time of diagnosis and denoted as A, B, C, D and E.

susceptible at the time of entry given the entry and diagnosis times of all 5 patients is computed by both the 2-body and 1-body mean field. In the 2-body mean field method, we compute with 2 patients and the infectivity contributed by the other three will be added into the background. We also want to test the effect of the proximity of the diagnosis times of the 2 patients (one of which is fixed as patient

E) used in the 2-body mean field method. We choose the second patient for the 2-body mean field method as D, C, B and A one by one, while treating the infectivity from others as background. Therefore, there are 4 estimations of E having latent infection at the time of entry, which are computed by the 2-body mean field method using DE, CE, BE and AE. The illustrations for the case with DE and CE are shown in Figure 5.5. For the 1-body mean field method and the naïve method, patients A - D are put into the background. Together with these two results, we have 6 total estimations for 27 experiments. The ROC curves are plotted and AUC are computed. The results are shown in Table 5.5. Here are the findings based on the results: 1) The values of the AUC of all the 2-body mean field method and the 1-body mean field method are greater than 0.5. This indicates the model has discriminating power; 2) The 1-body mean field method has the best average AUC as 0.7236 and the 2-body mean field method using D and E has the second best average AUC as 0.7209; 3) Naïve method has the worst average AUC as 0.5539, just slightly better than random guess, which has a AUC of 0.5;

In order to test the difference among the 2-body mean field method (different ordering), approximation and naïve method, we perform paired t-test on the experiment results. We followed a standard paired t-test [56]:

- Let R_1, R_2 be $N \times 1$ vectors representing the N results generated by Model 1 and and Model 2 respectively.
- The null hypothesis, $\mathcal{H}_0$: Model 2 dose not generate better results than Model 1.
- t-statistics: $t = \dfrac{\mu_X}{\sigma_X}$, where $X = R_2 - R_1$, μ_X is the sample mean of X and σ_X is the standard deviation.
- $\mathcal{H}_0$ is rejected if $t > \tilde{t}_{1-c,v}$. $\tilde{t}_{1-c,v}$ is the critical value with degree of freedom v. Here $v = 26$ and c is chosen to be 0.05: $\tilde{t}_{0.95,26} = 1.7056$.

Here we have 27 experiments results (AUC) for 2-body mean field method with 4 different patient orderings, 1-body mean field method and the naïve method, a total of 6 different models. The t-statistics of all the pair of models are shown in

Table 5.5: The results of the AUC of the ROC curves computed on 27 experiments with different methods. The first 4 columns are the results of the 2-body mean field method. The two character indicates which two patients are used in the computation. The 5th columns shows the results of the 1-body mean field method and the last columns shows the results for the naïve method. The mean and standard deviation (STD) of the AUC across all 27 experiments are shown in the last two rows.

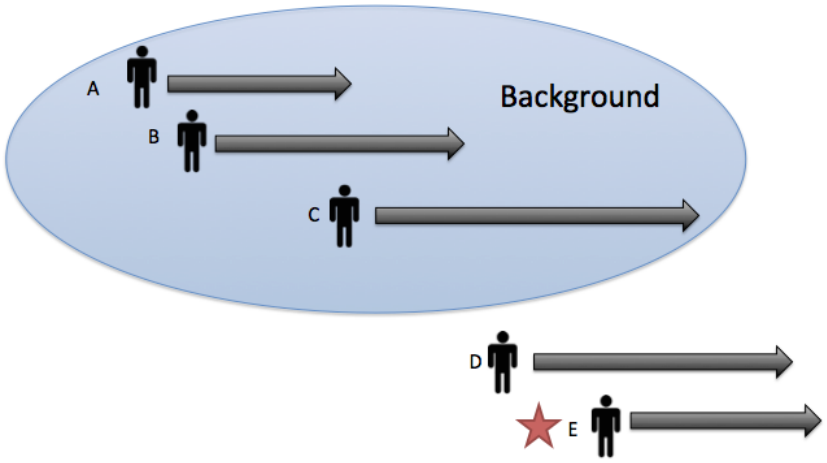
Exp No.	DE	CE	BE	AE	1-Body	Naïve
1	0.8471	0.8434	0.8337	0.8471	0.8475	0.5386
2	0.7708	0.7804	0.7891	0.7797	0.7771	0.6070
3	0.6696	0.6633	0.6565	0.6426	0.6665	0.6262
4	0.7999	0.8043	0.7931	0.7934	0.8048	0.5110
5	0.7475	0.7524	0.7353	0.7243	0.7455	0.5392
6	0.7436	0.7197	0.7272	0.7417	0.7509	0.5740
7	0.8151	0.7997	0.8032	0.8123	0.8204	0.5139
8	0.7526	0.7352	0.7375	0.7539	0.7658	0.5239
9	0.7276	0.7120	0.7121	0.7187	0.7367	0.5827
10	0.7337	0.7222	0.7154	0.7294	0.7316	0.5052
11	0.7119	0.7170	0.6928	0.6921	0.7096	0.5734
12	0.6828	0.6780	0.6796	0.6841	0.6792	0.5742
13	0.7928	0.7830	0.7813	0.7920	0.7951	0.4844
14	0.7659	0.7558	0.7626	0.7529	0.7688	0.5788
15	0.6933	0.6807	0.6786	0.6848	0.6980	0.5966
16	0.7495	0.7489	0.7468	0.7579	0.7621	0.5453
17	0.7429	0.7325	0.7377	0.7420	0.7480	0.5249
18	0.6472	0.6453	0.6419	0.6373	0.6532	0.5867
19	0.7596	0.7552	0.7632	0.7646	0.7608	0.5065
20	0.7166	0.7174	0.7093	0.7102	0.7132	0.5311
21	0.5943	0.5988	0.5971	0.5960	0.5965	0.6552
22	0.6942	0.7031	0.7002	0.6939	0.6960	0.5007
23	0.7065	0.7101	0.7151	0.7115	0.7139	0.5568
24	0.5751	0.5760	0.5792	0.5792	0.5783	0.5870
25	0.7264	0.7157	0.7216	0.7254	0.7247	0.5347
26	0.6936	0.6889	0.6876	0.6871	0.6914	0.5543
27	0.6028	0.6000	0.5967	0.5927	0.6013	0.5427
Mean	0.7209	0.7163	0.7146	0.7165	0.7236	0.5539
STD	0.0648	0.0633	0.0633	0.0664	0.0660	0.0411

Table 5.6. For example, the entry with row label "BE" and columns label "Naïve" is the t-statistics computed based on $X_{BE} - X_{\text{Naïve}}$, where X_{BE} is the AUC values of the 2-body mean field method using patient B and E and $X_{\text{Naïve}}$ is the AUC values of the Naïve method. Here $t = 1.7472$ is greater than the critical value: 1.7056. Therefore, the null hypothesis, "the performance of 2-body mean field method with B, E is not better than the naïve method", is rejected.

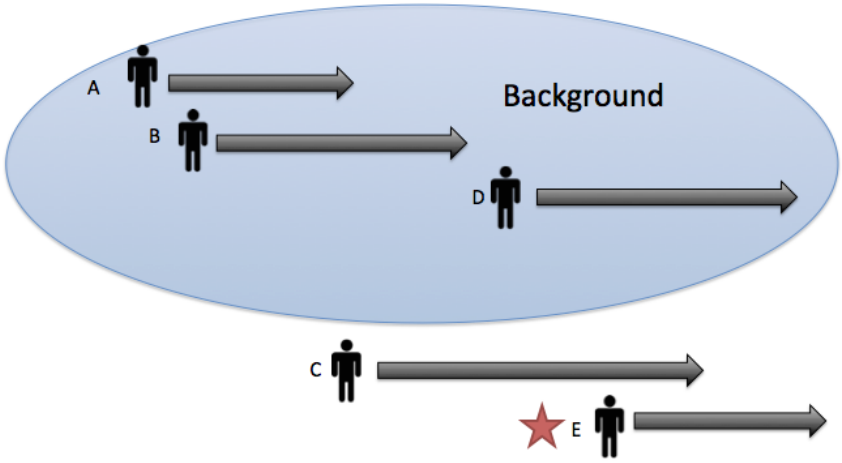
Table 5.6: T-statistics of the paired t-test for pairs of experiments. "Approx" means the 1-body mean field method.

	DE	CE	BE	AE	1-Body	Naïve
DE	0	0.5364	0.6884	0.4829	-0.5761	1.7815
CE	-0.5364	0	0.2113	-0.0238	-0.6786	1.7620
BE	-0.6884	-0.2113	0	-0.2332	-0.9224	1.7472
AE	-0.4829	0.0238	0.2332	0	-0.8863	1.7018
1-Body	0.5761	0.6786	0.9224	0.8863	0	1.7956
Naïve	-1.7815	-1.7620	-1.7472	-1.7018	-1.7956	0

Based on the results, the 2-body mean field method with patients DE, CE, BE and 1-body mean field method have significantly better performance than naïve method. The 1-body mean field methods are slightly better than the 2-body mean field method with all 4 patient orderings, but these differences are not statistically significant. One should note that the 1-body mean field method is much more computationally efficient than the 2-body mean field method. One run of the 2-body mean field method with 1000 clusters of size 5 take an approximately 20 hours, while for the 1-body mean field method it takes less than 1 minute (Computer: Mac OSX, 2.26 GHz Intel Core 2 Duo). Therefore, the 1-body mean field method is preferred in our model and the following experiments will be computed only using this method.



(a) DE



(b) CE

Figure 5.5: The figures show 2 the patient orderings for the 2-body mean field method. Figure (a) shows an illustration of computing with patient D and E (with E as the target) and A, B and C are in the background. Figure (b) shows computing with C and E (with E as the target) and A, B and D are in the background.

Different Cluster Sizes: We also test the effect of the cluster size. Again, we use the patient E as our target patient. We compute the conditional probability of E being susceptible at entry with only D and E in the cluster. Then we increase the size of the clusters by adding C, B and A one by one. i.e. "CDE" represents the setting where C,D and E are in the cluster. We use the 1-body mean field method with "DE", "CDE", "BCDE" and "ABCDE". The conditional probability is recomputed each time. The resulting AUC are shown in Table 5.7. The cluster of size 5 (last column, "ABCDE") on average has the best performance. This indicates that the model is more accurate with more information about the cluster. A paired t-test is perform for these four models and the results are shown in Table 5.8. The format of Table 5.8 is the same as the previous section. Based on these results, adding in more patients into the clusters always generate better AUC. However, the improvements of adding more patients are not statistically significant (all the t-statistics are smaller than $\tilde{t}_{0.95,26} = 1.7056$).

Table 5.8: T-statistics of the paired t-test for pairs of experiments.

	DE	CDE	BCDE	ABCDE
DE	0	-0.8317	-1.1166	-0.8426
CDE	0.8317	0	-0.7728	-0.5597
BCDE	1.1166	0.7728	0	-0.0622
ABCDE	0.8426	0.5597	0.0622	0

Table 5.7: The results of the AUC of the ROC curves computed on 27 experiments with different cluster sizes using the 1-body mean field method. E is our target patient. The characters represents the patient who are added in the cluster. e.g. "CDE" represents the setting where C,D and E are in the cluster. The mean and standard deviation (STD) of the AUC across 27 experiments are shown the last two rows.

Exp No.	DE	CDE	BCDE	ABCDE
1	0.8219	0.8301	0.8454	0.8475
2	0.7917	0.7852	0.7898	0.7771
3	0.6409	0.6534	0.6617	0.6665
4	0.8096	0.8084	0.8051	0.8048
5	0.7308	0.7382	0.7416	0.7455
6	0.7473	0.7473	0.7498	0.7509
7	0.8083	0.8133	0.8174	0.8204
8	0.7495	0.7554	0.7624	0.7658
9	0.7302	0.7356	0.7374	0.7367
10	0.7217	0.7294	0.7294	0.7316
11	0.7142	0.7079	0.7150	0.7096
12	0.6731	0.6776	0.6814	0.6792
13	0.7914	0.7951	0.7969	0.7951
14	0.7540	0.7617	0.7670	0.7688
15	0.6826	0.6886	0.6931	0.6980
16	0.7476	0.7566	0.7597	0.7621
17	0.7375	0.7421	0.7489	0.7480
18	0.6423	0.6469	0.6497	0.6532
19	0.7532	0.7687	0.7632	0.7608
20	0.7114	0.7151	0.7179	0.7132
21	0.5964	0.5951	0.5960	0.5965
22	0.6903	0.6900	0.6925	0.6960
23	0.7074	0.7100	0.7099	0.7139
24	0.5796	0.5775	0.5778	0.5783
25	0.7190	0.7271	0.7280	0.7247
26	0.6916	0.6951	0.6939	0.6914
27	0.5912	0.5965	0.5995	0.6013
Mean	0.7161	0.7203	0.7233	0.7236
STD	0.0652	0.0658	0.0666	0.0660

Extra Patients after Diagnosis: After testing the effects of patient ordering and cluster size, we are also interested in the effect of adding in patients who are diagnosed after the target patient. Instead of patient E, we choose patient C as our target. The conditional probability of C being susceptible at entry is computed using the 1-body mean field method under two different settings: 1) patient A, B in the cluster. This case is denoted as "ABC"; 2) in addition to A and B, adding in patient D and E. This case is denoted as "ABDEC". The AUC of the ROC curves are shown in Table 5.9. 25/27 experiments generate better performance, as measured by AUC, after incorporating the information of the patients who are diagnosed after the target patient. A paired t-test is performed with the null hypothesis that 1-body mean field method with "ABDEC" dose not have better performance than "ABC". The t-statistics is 1.2725. Although adding extra patients after the diagnosis time of the target patient improves the performance of the model, the improvement is not statistically significant.

Table 5.9: The AUC results of all the experiments of two configurations. One is setting target patient as C with A and B in the cluster (ABC). The other also uses C as target and adds A, B, D and in the cluster. Note that D and E are patients who are diagnosed after C's diagnosis time. ("STD" represents standard deviation)

Exp No.	ABC	ABDEC	Exp No.	ABC	ABDEC
1	0.8067	0.9194	15	0.6759	0.6619
2	0.7530	0.8368	16	0.7477	0.8454
3	0.7049	0.7642	17	0.7188	0.7588
4	0.8182	0.8946	18	0.6418	0.6670
5	0.7745	0.8244	19	0.7538	0.8485
6	0.6963	0.7441	20	0.6923	0.7121
7	0.7777	0.8697	21	0.6565	0.6666
8	0.7637	0.7941	22	0.7522	0.8233
9	0.7097	0.7338	23	0.6854	0.6995
10	0.7176	0.8756	24	0.6191	0.6388
11	0.7249	0.7581	25	0.6921	0.7634
12	0.7208	0.7065	26	0.6495	0.6922
13	0.7572	0.8526	27	0.6141	0.6170
14	0.6859	0.7565			
			Mean	0.7152	0.7676
			STD	0.0501	0.0839

Sensitivity on Parameters: As discussed in Chapter 4, estimating the parameters from the data is impractical. In this section, we test the sensitivity of the model performance to the parameters. We would like to choose 3 parameter settings, under which our model performs the best, average and worst. They are setting 1(best), 9(average) and 24(worst). Similarly, we set E as target patient, while A, B, C and D are in the cluster. For each of data sets 1, 9 and 24, the model is run with all parameter settings. For example, for data set simulated with parameter setting 9, model is run with settings 1-27. Note that here we would like to test the performance of our model while the parameters used are not accurate. We use the 1-body mean field method and the AUC for the experiments are computed.

Let $U_{i,j}$ be the value of AUC of the data simulated with the parameter setting i, and computed with setting j. For example, we use parameter setting 9 to simulate data. Then this data is computed with the 1-body mean field method using the same parameter setting. The AUC generated by this experiment is denoted as $U_{9,9}$. In addition to compute with parameter setting 9, the other 26 parameter settings are also used in the 1-body mean field method. These experiments will generate $U_{9,i}$ with $i = 1, 2, ..., 27; i \neq 9$. In other words, $U_{i,i}$ is the AUC computed with the "true parameters" and $U_{i,j;j=1,2,...,27;j\neq i}$ are the AUC computed with the "wrong parameters". Let μ_i^{AUC} be the average value and σ_i^{AUC} be the standard deviation of $\{U_{i,j}\}_{j=1,2,...,27}$. The results are presented in the following:

- $U_{1,1} = 0.8475$, $\mu_1^{AUC} = 0.8437$, $\sigma_1^{AUC} = 0.0038$
- $U_{9,9} = 0.7367$, $\mu_9^{AUC} = 0.7324$, $\sigma_9^{AUC} = 0.0036$
- $U_{24,24} = 0.5783$, $\mu_{24}^{AUC} = 0.5761$, $\sigma_{24}^{AUC} = 0.0016$

A boxplot of $\{U_{i,j}\}_{j=1,2,...,27;i\neq j}$ for i = 1, 9 and 24 are shown in Figure 5.6 ($U_{i,i}$ is plotted as black circle). As shown in the figure, the values of AUC are insensitive to the parameters.

Recall that the initial status of the 1000 patient E is either latent or susceptible. The conditional probability of E being susceptible, v_s, is computed by our model and is used as the score to determine whether E is considered to be susceptible at entry or not. We set a threshold s with initial value 1 and gradually reduce it to 0.

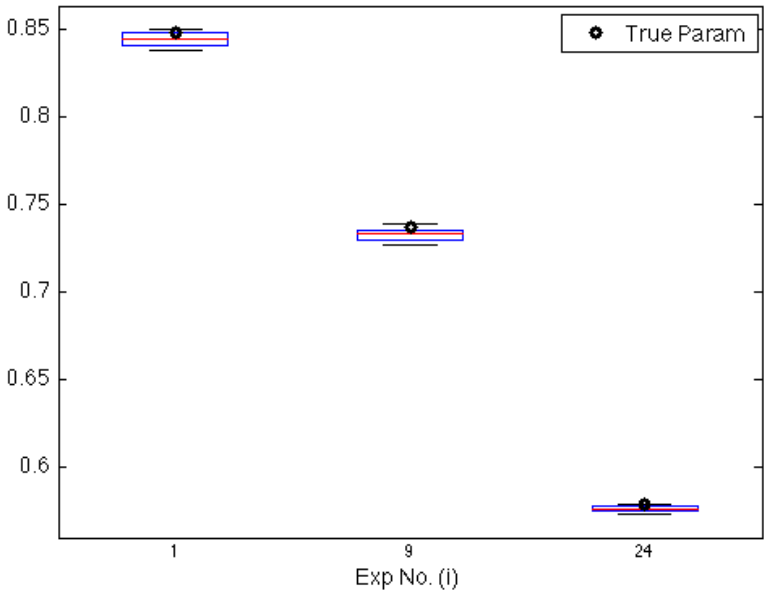


Figure 5.6: The boxplot of $\{U_{i,j}\}_{j=1,2,...,27}$ for $i = 1$, 9 and 24. $U_{i,j}$ is the value of AUC of the model results using data set simulated with parameter setting i, but computed with setting j. The AUC using the true parameters, i.e. $U^i_i, i = 1, 9, 24$ are plotted in black circles. In all experiments, we use the 1-body mean field method while setting E as target patient and using all 5 patients in the cluster.

Each step, the patients with $v_s \geq s$ are classified as being susceptible at entry. The threshold s starts at 1 and gradually decreases; we stop once we have more than 50% True Positive Rate (TPR), i.e. patients who entered susceptible and are classified as susceptible based on their v_s. The values of False Positive Rate (FPR) are recorded. Let $\text{TPR}_{i,j}$ ($\text{FPR}_{i,j}$) be the True Positive Rate (False Positive Rate) of the results which used data set simulated with parameter setting i and is computed with setting j. The scatter plot of the $\text{TPR}_{i,j}$ versus $\text{FPR}_{i,j}$ for $i = 1, 9, 24$ and $j = 1, 2, ..., 27$ are shown in Figure 5.7. The sub-figure on the left hand sides shows the result computed using the simulated data with parameter setting 1, with which our model has the best performance. It is shown that at the threshold values where our model has a 50% true positive rate, the false positive rate is approximately 11% false positive rate. In other words, when our model successfully identify 50% of the true

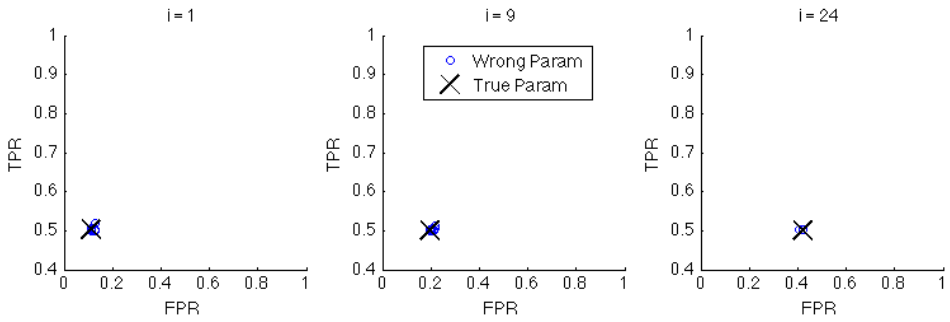


Figure 5.7: The scatter plot of $\text{TPR}_{i,j}$ (first one greater than 50%) versus $\text{FPR}_{i,j}$ for $i = 1, 9, 24$. The crosses represent the values plotted with $\text{TPR}_{i,i}$ versus $\text{FPR}_{i,i}$.

positive cases (the target patient entered susceptible and is identified as so), it only includes 11% of the false positive cases (the target patient entered susceptible and is identified as latent). The middle figure shows the results computed using parameter setting 9, under which our model has an average AUC. At the threshold value where our model has 50% true positive rate, the false positive rate is approximately 20%. The right most figure displays the results computed using simulated data with parameter setting 24, under which our model has the worst performance. Here at the threshold value where our model has 50% true positive rate, the false positive rate is about 40%; it is still better than random guess. Within each sub-figure, it is shown that the values $\text{TPR}_{i,j}$ and $\text{FPR}_{i,j}$ using different parameter settings are clustered closely with each other. This indicates that our model is consistent regardless of which parameter setting it used. Once again, we have shown that our model is insensitive to parameters.

The consistency is a result of the characteristics of our model. Before discussing this, a few notations need to be introduced. For each cluster, the conditional probability of patient E being susceptible given the entry and diagnosis times of all the 5 patients is computed. Let $y_{i,j}^{(k)}$ be the value of this conditional probability of the k^{th} cluster in the data set which is simulated with parameter settings i and computed with the 1-body mean field method using setting j. As shown in Figure 5.6, although using the wrong parameters, the AUC are consistent with the one computed with the true parameters. This is because the shape of ROC curves only depends on the ranking of the clusters based on the order of $y_{i,j}^{(k)}$. For example, for

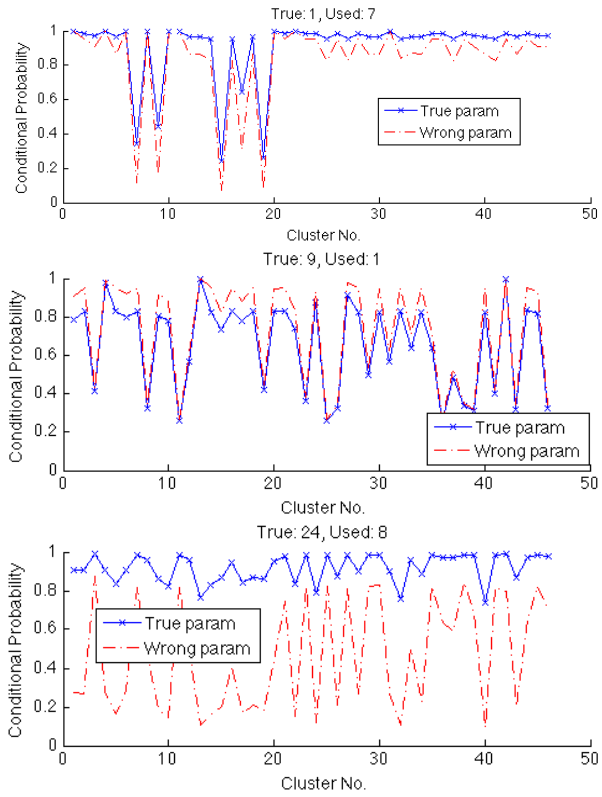


Figure 5.8: Top: Data set simulated with setting 1, computed with setting 7; Middle: Data set simulated with setting 9, computed with setting 1; Bottom: Data set simulated with setting 24, computed with setting 8. For the conditional probabilities in each plot, while the values are different with different parameters in the computation, the relative positions remain approximately the same. For example, when we investigate the 3rd and 4th values in the bottom plot, we have $y^{(3)}_{24,24} \geq y^{(4)}_{24,24}$ and $y^{(3)}_{24,8} \geq y^{(4)}_{24,8}$.

any pair of m, n in 1-1000, if $y^{(m)}_{i,p} \leq y^{(n)}_{i,p}$ implies $y^{(m)}_{i,q} \leq y^{(n)}_{i,q}$, the shape of the ROC curve of the model using parameter setting p will be exactly the same as the one using setting q. A sample of 45 values $y^{(k)}_{i,j}$ for three different setting of i and j are plotted in Figure 5.8. In our case, for data set simulated using setting i, the ranking of the clusters are approximately the same regardless which parameter settings are used in the computation. Therefore the ROC curves and the corresponding AUC of our model are consistent even when we used different parameter settings.

Domestic Bath: So far we have been using our model without the domestic bath. We are going to test our model with the domestic bath. The probability of the target patient being infected by the domestic bath is β_d. Similarly, the new experiments will have 5 patients in the cluster with E as the target. We choose experiments with parameter settings: 1, 9 and 24, with the same reason as previous sections.

The patient clusters are simulated with the same procedure as in Algorithm 1. The simulated data contain the entry and diagnosis times of TB patients; there is a percentage of patient who were susceptible at the time of entry (these patients were infected after entry). This percentage is denoted as p_s. p_s will increase as β_d increases, since it is more likely for a susceptible person to be infected. We choose the two different values of β_d: one will increase p_s by approximately 10% the other will increase p_s by approximately 20%, compared to the simulated data without domestic bath ($\beta_d = 0$). The purpose of doing this is to evaluate the performance of the model with influence of the domestic background with two different intensities. The values of β_d chosen for each experiment and the corresponding p_s and AUC are shown in Table 5.10.

Table 5.10: Experiments with parameter settings 1,9 and 24 are simulated again with two different values of background infectivity, β_d. The values of β_d are chosen to increase the original percentage of susceptible patients, p_s, by approximately 10% and 20%. The format of table entries is: "AUC, (β_d, p_s)".

Exp No.	$\beta_d = 0$	increase β_d	increase β_d again
1	0.8475, (0, 6.38%)	0.8005, ($3e^{-7}$, 7.10%,)	0.7560 , ($5e^{-7}$, 8.12%)
9	0.7367, (0, 34.66%)	0.7233, ($4e^{-6}$, 37.74%)	0.6797, ($8e^{-6}$, 40.96%)
24	0.5783 (0, 14.84%)	0.6113, ($4e^{-6}$, 16.58%)	0.6122, ($1.80e^{-5}$, 18.24%)

In the cases where our model has good performances, i.e. experiments 1 and 9, the presence of the domestic bath infectivity makes the performances decay. For experiment 1, our model has AUC of 0.8475 when $\beta_d = 0$ and $p_s = 6.38\%$; it drops to 0.7560 when $\beta_d = 5 \times 10^{-7}$ and $p_s = 8.12\%$. On the other hand, in experiment 24, where our model performs even worse than the naïve method, adding

in the domestic bath improves our model. Our model has AUC of 0.5783 with $p_s = 14.84\%$ to 0.6122 with $p_s = 18.24\%$. For each parameter settings, we compute the conditional probability of the target patient in each cluster being susceptible at entry at different β_d. We plot the distribution of the conditional probabilities of the target patients who actually entered susceptible, compare it to the distribution of the conditional probabilities of those who actually entered with latent infection. The plots are presented in Figure 5.9. For results with parameter settings 1 and 9, adding the domestic bath infectivity has negative impact on the model performance and the two distributions become less "distinguishable" as β_d increase. On the other hand, the two distributions for the experiments with parameter setting 24 are close to each other even without domestic bath infectivity. That is why the AUC for this case is merely 0.5783, indicating it is difficult to distinguish latent person and the susceptible ones based on the conditional probabilities our model computed. Increasing β_d changes the data and makes the two distribution more distinct and thus easier to distinguish latent and susceptible persons (better AUC).

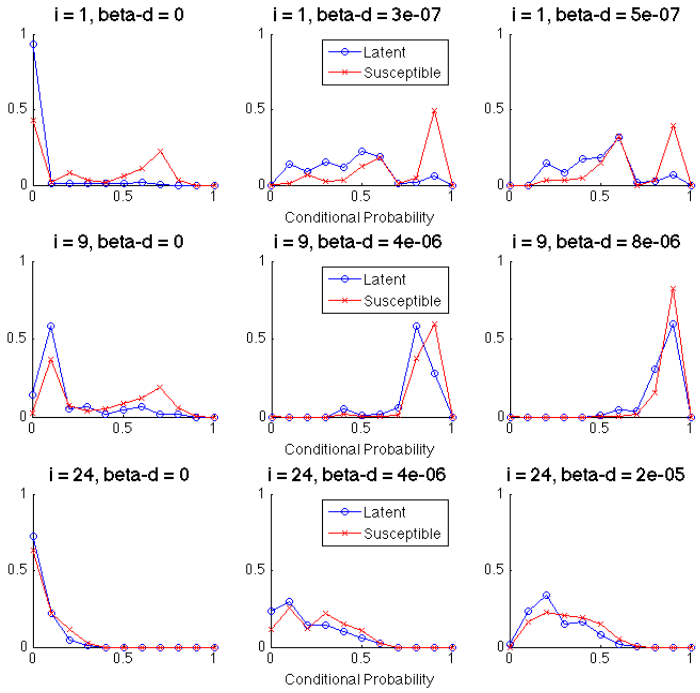


Figure 5.9: The conditional probabilities of the target patients being susceptible at entry were computed. Each sub-figure presents a plot of two distributions of the values of these computed conditional probabilities: 1) red cross represents the distribution of the conditional probabilities of those who were actually susceptible at entry; 2) blue circle represents the distribution of the conditional probabilities of those who were actually latent at entry. From top to bottom, row 1 displays the results with parameter setting 1, row 2 displays the results with parameter setting 9 and row 3 displays the results with parameter setting 24.

Hidden Active Foreign Bath: We also tested the effect of the hidden active foreign bath. Assume we have a cluster of size n and the latest diagnosis time is t_n. The hidden active foreign bath represents the foreign-born patients who have not been diagnosed by t_n. We test this by adding in an special foreign-born patient into the background, whose infectivity at t_n equals to the infectivity contributed by an average number of active patients in the population. This infectivity takes a geometric decay with rate γ stepping backward away from t_n. It is shown that the effect of the hidden active foreign bath is negligible for our model's performance.

5.4 Application to the New York City Data

Finally, our model is applied to the real data from New York City. Based on our analysis, our model has discriminating power and is insensitive to parameters. We propose to use our model in the following way: 1) Use one set of reasonable parameters to compute; 2) Rank the cluster according to their target patient's probabilities being susceptible at entry; 3) Choose the top quantile of clusters, where recent TB transmissions are most likely to happen and investigate those clusters in detail. Instead of determining whether a patient was susceptible or not at entry for all the patient clusters, the model raises a flag at the most suspicious ones so that the healthcare officials could allocate the limited resources to investigate these ones.

Inferring parameter values from the data is impractical, as discussed in Chapter 4. We set a high, median, low for parameters α, Γ and π. For applying our model to the real data, we choose the median values for these three parameters. This choice is somewhat arbitrary but we can do this because our model is shown to be insensitive to α, γ and π. We showed this by experimenting our model with different combinations of the three values (low-median-high) for these variables. For the NYC data, we choose the values to b: $\alpha = 1 \times 10^{-4}$, $\Gamma = 1$, $\pi = 0.07$. For γ and δ, β_d, we also choose the values used in the experiments: $\gamma = 0.3333$, $\delta = 0.05$ and $\beta_d = 0$. The total foreign-born population in NYC, N_F, is set to be 3,000,000. Since the probability to be infected in a particular month is going to be diluted by the total foreign-born population, the computed conditional probability of the target patient being susceptible at entry will be very small (around 1×10^{-4}). We can still

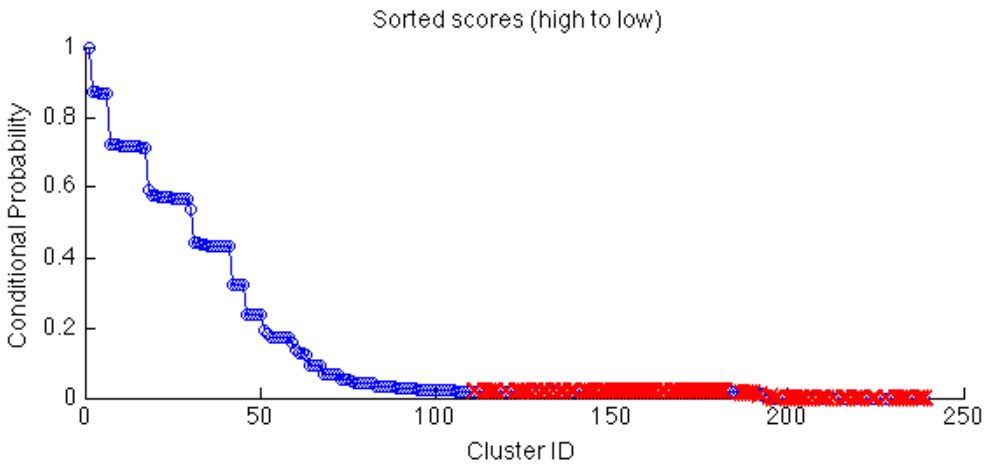


Figure 5.10: For the 239 clusters with size 2 among the NYC data, the scores of the target patients being susceptible at entry are computed by the 1-body mean field method and plotted in a descending order. For the clusters in which the two patients' diagnosis times are more than 2 years apart, we plot a red cross.

rank the patients according to these probabilities, but they are inconvenient to read. Since we are more interested in finding the clusters in which the target patients have relatively high conditional probabilities of being susceptible at entry, we normalize these conditional probabilities by its maximum value. This means that the cluster with highest conditional probability will have a score of 1 and all the scores are positive. The patient who was diagnosed last in the cluster is set as the target. We use the 1-body mean field method to compute the conditional probability of the target patient being susceptible at entry and then normalize it to produce a score. For the clusters with target patients having high scores, transmissions are likely to have happened. Therefore, these clusters are said to be "suspicious".

Among all the clusters in the NYC data set, we first choose the clusters with size 2 and there are 239 such clusters. The 239 scores of the target patients are sorted from high to low and shown in Figure 5.10. Note that the current method used by the healthcare workers to identify transmission is to see if the diagnosis times of two patients, given they share the common TB genotypes, are within 2 years [57]. For the clusters in which the two patients' diagnosis times are more than 2 years apart, we plot a red cross. As shown in Figure 5.10, the current method

[57] of identifying transmission provides a coarse result: nearly half of clusters are classified as containing transmissions. On the other hand, our model provides a much more refined result: allowing us to pick an arbitrary fraction of the clusters, in which transmissions are most likely. The entry and diagnosis times of the 20 most suspicious (highest conditional probability) clusters with the computed scores are show in Table 5.11. Each cluster is classified by its Spoligotype ID, which is a ID assigned to each Spoligotype, and RFLP. A patient may have contact with another patient within the cluster. This contact is referred as an epidemiological link (epi link). NYC Bureau of TB control has perform contact investigation on the clusters with the last case reported in 2007 and some of these clusters are shown to have epi links [58]. It turns out the target patients in the clusters with epi links have higher scores in our model. Based on the report from NYC Bureau of TB control [58], there is one cluster with size 2 which has epi links and it is in our data set: the cluster has Spoligotype ID: "S01800", RFLP: "IA", (referred as "S01800 IA"). The conditional probability of the target patient in this cluster being susceptible ranks the 4th among the 239. There is another cluster,"S00210, GD318", in which both patients lived in the same neighborhood and from the same country; the investigation of the epi link were in progress as reported in 2008. This cluster also has a high ranking: 12/239. Based on the same procedure, the clusters of size 3 (74 such clusters) are chosen. The patients who were diagnosed last are set as the target patient. The scores are plotted in a descending order in Figure 5.11. For each cluster, let $I^{(a)}$ be the patient whose diagnosis time is the closest to the target patient's. For the clusters in which $I^{(a)}$'s and the target patient's diagnosis times are more than 2 years apart, a red cross is plotted. The information of the top 15 suspicious clusters are shown in Table 5.12. Two clusters in our data set are reported to have epi links in the report [58]: "S00540 BM45" and "S00034 W966". The target patients in these two clusters again have high scores in our model: 8^{th} and 11^{th} in 74 clusters.

5.5 Conclusion

In this section, we tested our model with the simulated data. The performance of our model is evaluated with an ROC curve. It is shown that our model has

Table 5.11: The information of 20 (out of 239) most suspicious clusters of size 2. The third columns shows the entry/diagnosis time (unit: month) of the patients within the cluster with the format: $[t_0^{(1)}, t_1^{(1)}; t_0^{(2)}, t_1^{(2)}]$. The scores of clusters are shown in the last column. Cluster "S01800 IA" has epi links. Patients in cluster "S00210 GD318" have close proximity; investigation was in progress.

SpoID	RFLP	Entry/Diagnosis Times	Score
S00282	HJ7	[1,2;1,2]	1.0000
S00437	AI212	[1,244;61,244]	0.7073
S01400	BW221	[1,24;14,24]	0.7032
S01800	**IA**	[1,307;293,307]	0.6971
S00476	CS54	[1,51;18,51]	0.6968
S01223	BA57	[1,103;86,103]	0.6962
S00034	DN	[52,291;1,292]	0.4792
S00050	BJ99	[1,256;46,257]	0.4753
S01095	BA90	[1,203;1,204]	0.4750
S00182	AX32	[1,157;8,158]	0.4724
S00034	DN5	[1,249;114,250]	0.4718
S00210	**GD318**	[1,209;82,210]	0.4714
S00671	NY	[10,123;1,124]	0.4712
S00009	C49	[13,102;1,103]	0.4702
S00682	GZ	[1,121;57,122]	0.4685
S00034	W338	[14,34;1,35]	0.4670
S00197	MX	[1,401;378,402]	0.4666
S00196	HH10	[594,687;1,689]	0.3348
S00074	CS77	[156,348;1,350]	0.3237
S00197	HR102	[245,337;1,339]	0.3233

discriminating power in identifying transmission versus latent reactivation. The model is shown to be insensitive to parameters by experimenting with different combinations of parameter values.

Finally, we applied our model to the New York City data [46]. For each target patient in the clusters, the conditional probability being susceptible at entry, which means there is a TB transmission, was computed by our model. The target patients in the clusters which contain epi links were assigned high scores by our model. Contact investigations need to be perform in order to detect epi link and better control TB spreading. Due to limited staff and resources, contact investigations can only be done for a fraction of the TB clusters [58]. Healthcare officials can use our

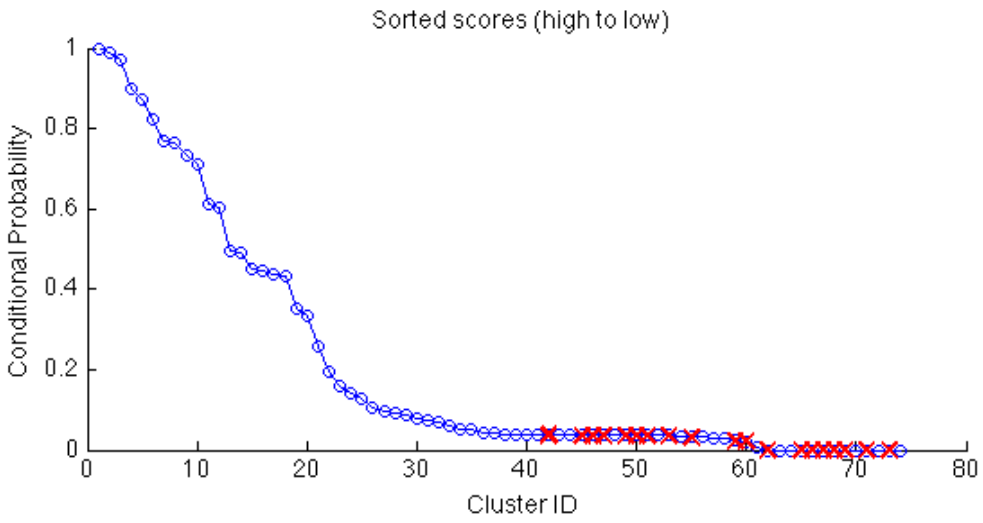


Figure 5.11: For the 164 clusters with size 3 among the NYC data, the scores of the clusters are computed by the 1-body mean field method and plotted in a descending order. Let $I^{(a)}$ be the patient whose diagnosis time is the closest to the target patient's. For the clusters in which $I^{(a)}$'s and the target patient's diagnosis times are more than 2 years apart, a red cross is plotted.

model to prioritize the cluster to investigate.

Table 5.12: The information of 15 (out of 74) most suspicious clusters of size 3. The third column shows the entry/diagnosis time (unit: month) of the patients within the cluster with the format: $[t_0^{(1)}, t_1^{(1)}; t_0^{(2)}, t_1^{(2)}; t_0^{(3)}, t_1^{(3)}]$. The scores of the clusters being susceptible at entry are shown in the last column. Clusters "S00540 BM45" and "S00034 W966" are shown to have epi links.

SpoID	RFLP	Entry/Diagnosis Times	Score
S00363	AI199	[88,120;1,122;51,122]	1.0000
S00002	LE11	[1,37;2,40;36,40]	0.9647
S00210	GD46	[1,95;52,95;76,96]	0.9197
S01141	BW321	[439,619;1,625;556,625]	0.7565
S00009	C4	[1,332;75,387;210,387]	0.7041
S00005	AX34	[1,153;136,153;132,155]	0.6179
S01776	AH5	[185,255;1,260;63,261]	0.5358
S00540	**BM45**	[134,173;1,178;57,179]	0.5317
S00200	W269	[162,246;1,254;189,255]	0.4866
S00034	W912	[1,183;77,215;198,216]	0.4603
S00034	**W966**	[1,513;475,514;356,517]	0.3541
S00026	MC10	[68,89;1,95;46,97]	0.3431
S00086	HP19	[1,618;638,638;624,639]	0.2545
S00074	CS13	[1,336;153,353;58,356]	0.2210
S00245	DK22	[106,124;10,154;1,157]	0.2176

CHAPTER 6
Conclusion and future work

This thesis studies TB from two aspects. The first part of the thesis investigates the evolution of the genome of the MTBC utilizing the information of the DNA finger printing data. The second part studies mathematical models for TB disease spread under the framework of small patient clusters.

6.1 MIRU Evolution

Conclusion: In this part, we utilize the database of DNA finger printing of the MTBC isolates to study the evolution of mycobacterial interspersed repetitive units (MIRU) in the MTBC genome. Based on spacer oligonucleotide types (spoligotypes), two rules are designed to infer the mutations and their direction of the isolates. A joint data set of 14,453 isolates gathered from United States Centers for Disease Control [28] and Institute Pasteur SITVIT [29] is examined to determine 41,604 of mutations.

We also investigated the dynamics of the mutations of the MIRU repeat numbers. We have found that different repeat numbers mutate differently. Small values (0-3) have high probability to increase while large numbers tend to decrease. It is found that it is more likely for a repeat number to change by smaller values than large ones. We also study the MIRU evolution by locus. We found that the mutation dynamics are different for different loci. Sticky values are defined as the number most of the repeat numbers tend to mutate to. This sticky value is different for different loci. A Markov chain model is built to investigate the future of the repeat number distribution, such as the stationary distribution and the convergence rate. Under the framework of Markov Chains, MIRU 24 is found to be the most stable locus, while MIRU 27 is the least stable one.

Future Work: This part of the study points out an interesting future research direction. In order to infer the mutation directions among isolates, we defined two

rules. These two rules take advantage of one of spoligotpes' characteristics: that it is easy for the DR region of the MTBC genome to lose a spacer, yet nearly impossible to gain one. These rules can be used to find the root of the MTBC isolates. Using the same Markov chain framework, we can estimate the length of time it takes for the MIRU repeat numbers to evolve from the root to the current distribution. It can also divide the isolate data into different lineages, therefore the age of different lineages can be investigated in terms of the Markov chain framework.

6.2 TB Spread

Conclusion: In the second part of this thesis, we develop mathematical models to understand the transmission dynamics of TB. With the help of DNA finger printing technologies, we clustered TB patients into small groups, with size 1 - 10. The patients in the same cluster share a genetically common TB strain. They have possibility to be infected by someone in the cluster. We built mathematical models to estimate the probability of whether an immigrant TB patient enter the country with latent infection, given the information of other immigrants in the cluster. Unlike most of the mathematical models for TB epidemiology, which study the transmission dynamics at the population level [27], we built a model which studies the spread at an individual level. We first built a detailed model to help us get insight into the problem. Then, we use a mean-field style approximation to simplify the computations. Based on the simulation data, our model is proven to have discriminating power. Although it is impractical to estimate the parameters from the data, it is found that our model is insensitive to parameters. The model can be used to raise alarms to the most suspicious clusters, where a recent TB transmission will most likely occur.

Future Work: The study in this part allows us to investigate the dynamics of TB transmission in small groups. We tried to model the exact transmission routes in as much detail as possible. While doing this give us insight into the true dynamics, it leads us to complicated models, introducing difficulties in computation and parameter estimation.

One promising direction is to simplify the model. (Please refer to the idea of Occam's razor [59]) With the experience of building detailed models of TB spread, we now understand the dynamics in a more detailed sense. Therefore, we can simplify the model without violating fundamental rules of the transmission dynamics. We use a bottom-up approach to build a probabilistic model to estimate the likelihood of transmissions. What we learned from our model is that the likelihood of transmissions within a cluster is related to the closeness of the diagnosis time of the patients in the cluster. One could use a top-down approach, starting from real data to find relationships between the structure of the entry/diagnosis times and epi links within a cluster.

REFERENCES

[1] World Health Organization. (2013) *Global Tuberculosis Report 2013*. [Online]. Available: http://apps.who.int/iris/bitstream/10665/91355/1/9789241564656_eng.pdf [Date Last Accessed, 12/01/2014].

[2] T. M. Daniel, "The history of tuberculosis," *Respiratory Medicine*, vol. 100, no. 11, pp. 1862–1870, Nov. 2006.

[3] I. Comas and S. Gagneux, "The past and future of tuberculosis research," *PLoS Pathogens*, vol. 5, no. 10, p. e1000600, Oct. 2009.

[4] World Health Organization. (2013, Oct) *Tuberculosis Fact Sheets No. 104*. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs104/en/ [Date Last Accessed, 12/01/2014].

[5] Centers for Disease Control and Prevention (CDC). (2011, Nov) *Tuberculosis(TB) Fact Sheets*. [Online]. Available: http://www.cdc.gov/tb/publications/factsheets/general/ltbiandactivetb.htm [Date Last Accessed, 12/01/2014].

[6] ——. *Basic TB Facts*. [Online]. Available: http://www.cdc.gov/tb/basics/ [Date Last Accessed, 12/01/2014].

[7] Z. Feng, C. Castillo-Chavez, and A. F. Capurro, "A model for tuberculosis with exogenous reinfection," *Theoretical Population Biol.*, vol. 57, no. 3, pp. 235–247, Jun. 2000.

[8] S. Lawn and A. Zumla, "Tuberculosis," *Lancet*, vol. 378, no. 9785, pp. 57–72, Jul. 2011.

[9] H. Guo and J. Wu, "Persistent high incidence of tuberculosis among immigrants in a low-incidence country: Impact of immigrants with early or late latency," *Math. Biosci. and Eng.*, vol. 8, no. 3, pp. 695–709, Jul. 2011.

[10] Centers for Disease Control and Prevention (CDC), "Trends in tuberculosis -United States, 2013," *Morbidity and Mortality Weekly Rep.*, vol. 63, no. 11, pp. 229–233, Mar. 2013.

[11] ——. *Reported Tuberculosis in the United States, 2013*. [Online]. Available: http://www.cdc.gov/tb/statistics/reports/2013/default.htm [Date Last Accessed, 12/01/2014].

[12] S. Gagneux and P. M. Small, "Global Phylogeography of *Mycobacterium tuberculosis* and implications for tuberculosis product development," *The Lancet Infectious Diseases*, vol. 7, no. 5, pp. 328–337, May. 2007.

[13] J. F. Reyes and M. M. Tanaka, "Mutation rates of spoligotypes and variable numbers of tandem repeat loci in *Mycobacterium tuberculosis*," *Infection, Genetics and Evolution*, vol. 10, no. 7, pp. 1046–1051, Oct. 2010.

[14] K. Ijaz, Z. Yang, H. S. Matthews, J. H. Bates, and M. D. Cave, "*M*ycobacterium Tuberculosis transmission between cluster members with similar fingerprint patterns," *Emerging Infectious Diseases*, vol. 8, no. 11, pp. 491–504, Nov. 2002.

[15] J. Glynn, E. Vyonycky, and P. Fine, "Influence of sampling on estimates of clustering and recent transmission of *m*ycobacterium tuberculosis derived from dna fingerprinting techniques," *Amer. J. of Epidemiology*, vol. 149, no. 4, pp. 366–371, Jun. 1999.

[16] C. Ozcaglar, "Algorithmic data fusion methods for tuberculosis," Ph.D. dissertation, Comp. Sci., RPI, Troy, NY, 2012.

[17] S. T. Cole *et al.*, "Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence," *Nature*, vol. 393, no. 6685, pp. 537–544, Jun. 1998.

[18] C. Sola, I. Filliol, E. Legrand, S. Lesjean, C. Locht, P. Supply, and N. Rastogi, "Genotyping of the *Mycobacterium tuberculosis* complex using MIRUS: Association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics," *Infection, Genetics and Evolution*, vol. 3, no. 2, pp. 125–133, Jul. 2003.

[19] P. F. Barnes and M. D. Cave, "Molecular epidemiology of tuberculosis," *New England J. of Medicine*, vol. 349, no. 12, pp. 1149–1156, Sep. 2003.

[20] A. Shabbeer, "Nonconvex nonsmooth optimization for bioinformatic," Ph.D. dissertation, Comp. Sci., RPI, Troy, NY, 2013.

[21] J. R. Driscoll, "Spoligotyping for molecular epidemiology of the *Mycobacterium tuberculosis* complex," in *Molecular Epidemiology of Microorganisms*. New York, NY, USA: Springer, 2009, vol. 551, pp. 117–128.

[22] R. Warren, E. Streicher, S. Sampson, G. Van Der Spuy, M. Richardson, D. Nguyen, M. Behr, T. Victor, and P. Van Helden, "Microevolution of the direct repeat region of *Mycobacterium tuberculosis*: Implications for interpretation of spoligotyping data," *J. of Clinical Microbiology*, vol. 40, no. 12, pp. 4457–4465, Dec. 2002.

[23] K. Kremer *et al.*, "Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: Interlaboratory study of discriminatory power and reproducibility," *J. of Clinical Microbiology*, vol. 37, no. 8, pp. 2607–2618, Aug. 1999.

[24] P. Supply *et al.*, "Proposal for standardization of optimized mycobacterial interspersed repetitive unit-variable-number tandem repeat typing of *Mycobacterium tuberculosis*," *J. of Clinical Microbiology*, vol. 44, no. 12, pp. 4498–4510, Dec. 2006.

[25] L. S. Cowan, L. Mosher, L. Diem, J. P. Massey, and J. T. Crawford, "Variable-number tandem repeat typing of *Mycobacterium tuberculosis* isolates with low copy numbers of IS6110 by using mycobacterial interspersed repetitive units," *J. of Clinical Microbiology*, vol. 40, no. 5, pp. 1592–1602, May. 2002.

[26] E. Mazars, S. Lesjean, A.-L. Banuls, M. Gilbert, V. Vincent, B. Gicquel, M. Tibayrenc, C. Locht, and P. Supply, "High-resolution minisatellite-based typing as a portable approach to global analysis of*Mycobacterium tuberculosis* molecular epidemiology," *Proc. of the Nat. Academy of Sci.*, vol. 98, no. 4, pp. 1901–1906, Feb. 2001.

[27] C. Ozcaglar, A. Shabbeer, S. L. Vandenberg, B. Yener, and K. P. Bennett, "Epidemiological models of *Mycobacterium tuberculosis* complex infections," *Math. Biosci.*, vol. 236, no. 2, pp. 77–96, Apr. 2012.

[28] M. Aminian, A. Shabbeer, and K. P. Bennett, "A conformal bayesian network for classificationof *Mycobacterium tuberculosis* complex lineages," *BMC Bioinformatics*, vol. 11, no. Suppl 3, p. S4, Apr. 2010.

[29] C. Ozcaglar, A. Shabbeer, S. Vandenberg, B. Yener, and K. P. Bennett, "Sublineage structure analysis of mycobacterium tuberculosis complex strains using multiple-biomarker tensors," *BMC Genomics*, vol. 12, no. Suppl 2, p. S1, Feb. 2011.

[30] S. I. Resnick, *Adventures in Stochastic Processes*. New York, NY, USA: Springer, 1992.

[31] G. Strang, *Introduction to Linear Algebra*. Cambridge, England: Cambridge Publication, 2003.

[32] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, England: Cambridge Univ. Press, 2008, vol. 1.

[33] J. F. Bromaghin, "Sample size determination for interval estimation of multinomial probabilities," *The Amer. Statist.*, vol. 47, no. 3, pp. 203–206, Aug. 1993.

[34] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Ann. of Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[35] A. Shabbeer, L. S. Cowan, C. Ozcaglar, N. Rastogi, S. L. Vandenberg, B. Yener, and K. P. Bennett, "TB-lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex," *Infection, Genetics and Evolution*, vol. 12, no. 4, pp. 789–797, Jun. 2012.

[36] C. Demay, B. Liens, T. Burguière, V. Hill, D. Couvin, J. Millet, I. Mokrousov, C. Sola, T. Zozio, and N. Rastogi, "Sitvitweb–a publicly available international multimarker database for studying *Mycobacterium tuberculosis* genetic diversity and molecular epidemiology," *Infection, Genetics and Evolution*, vol. 12, no. 4, pp. 755–766, Jun. 2012.

[37] Y.-J. Sun, R. Bellamy, A. S. Lee, S. T. Ng, S. Ravindran, S.-Y. Wong, C. Locht, P. Supply, and N. I. Paton, "Use of mycobacterial interspersed repetitive unit-variable-number tandem repeat typing to examine genetic diversity of*Mycobacterium tuberculosis* in Singapore," *J. of Clinical Microbiology*, vol. 42, no. 5, pp. 1986–1993, May. 2004.

[38] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: John Wiley & Sons, 2012.

[39] S. M. Blower, A. R. McLean, T. C. Porco, P. M. Small, P. C. Hopewell, M. A. Sanchez, and A. R. Moss, "The intrinsic transmission dynamics of tuberculosis epidemics," *Nature Medicine*, vol. 1, no. 8, pp. 815–821, Aug. 1995.

[40] C. Castillo-Chavez and B. Song, "Dynamical models of tuberculosis and their applications," *Math. Biosci. Eng.*, vol. 1, no. 2, pp. 361–404, Sep. 2004.

[41] P. Roche *et al.*, "Tuberculosis notifications in Australia, 2005," *Communicable Diseases Intell. Quart. Rep.*, vol. 31, no. 1, pp. 71–80, Mar. 2007.

[42] A. E. Fanning, "Globalization of tuberculosis," *CMAJ: Canadian Medical Assoc. J.*, vol. 158, no. 5, p. 611, Mar. 1998.

[43] E. Geng, B. Kreiswirth, C. Driver, J. Li, J. Burzynski, P. DellaLatta, A. LaPaz, and N. W. Schluger, "Changes in the transmission of tuberculosis in New York City from 1990 to 1999," *New England J. of Medicine*, vol. 346, no. 19, pp. 1453–1458, May. 2002.

[44] K. P. Cain, S. R. Benoit, C. A. Winston, and W. R. Mac Kenzie, "Tuberculosis among foreign-born persons in the United States," *Jama*, vol. 300, no. 4, pp. 405–412, Jul. 2008.

[45] United States Census Bureau. (2010) *The Foreign-Born Population in the United States: 2010*. [Online]. Available:

http://www.census.gov/prod/2012pubs/acs-19.pdf [Date Last Accessed, 12/01/2014].

[46] I. Vitol, J. Driscoll, B. Kreiswirth, N. Kurepina, and K. P. Bennett, "Identifying mycobacterium tuberculosis complex strain families using spoligotypes," *Infection, Genetics and Evolution*, vol. 6, no. 6, pp. 491–504, Nov. 2006.

[47] K. M. Ramachandran and C. P. Tsokos, *Mathematical Statistics With Applications*. Waltham, MA,USA: Academic Press, 2009.

[48] T. Fawcett, "An introduction to ROC Analysis," *Pattern Recognition Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[49] T. Chard, "REVIEW: Pregnancy tests: a review," *Human Reproduction*, vol. 7, no. 5, pp. 701–710, May. 1992.

[50] New York City Department of City Planning - Population Division. (2013, Dec) *The Newest New Yorkers - Characteristics of the City's Foreign-born Population*. [Online]. Available: http://www.nyc.gov/html/dcp/pdf/census/nny2013/nny_2013.pdf [Date Last Accessed, 12/01/2014].

[51] K. M. Shea *et al.*, "Estimated rate of reactivation of latent tuberculosis infection in the United States, overall and by population subgroup," *Amer. J. of Epidemiology*, vol. 179, no. 2, p. 246, Jan. 2013.

[52] C. R. Horsburgh Jr, M. O'Donnell, S. Chamblee, J. L. Moreland, J. Johnson, B. J. Marsh, M. Narita, L. S. Johnson, and C. F. von Reyn, "Revisiting rates of reactivation tuberculosis: a population-based approach," *Amer. J. of Respiratory and Critical Care Medicine*, vol. 182, no. 3, pp. 420–425, Apr. 2010.

[53] P. M. Ricks, K. P. Cain, J. E. Oeltmann, J. S. Kammerer, and P. K. Moonan, "Estimating the burden of tuberculosis among foreign-born persons acquired prior to entering the US, 2005–2009," *PLoS One*, vol. 6, no. 11, p. e27405, Nov. 2011.

[54] N. D. Walter, J. Painter, M. Parker, P. Lowenthal, J. Flood, Y. Fu, R. Asis, and R. Reves, "Persistent latent tuberculosis reactivation risk in United States immigrants," *Amer. J. of Respiratory and Critical Care Medicine*, vol. 189, no. 1, pp. 88–95, Jan. 2014.

[55] Y. Liu, J. A. Painter, D. L. Posey, K. P. Cain, M. S. Weinberg, S. A. Maloney, L. S. Ortega, and M. S. Cetron, "Estimating the impact of newly arrived foreign-born persons on tuberculosis in the United States," *PloS One*, vol. 7, no. 2, p. e32158, Feb. 2012.

[56] H. Motulsky, *Intuitive Biostatistics: a Nonmathematical Guide to Statistical Thinking*. Oxford, England: Oxford Univ. Press, 2013.

[57] Centers for Disease Control and Prevention (CDC). *Guide to the Application of Genotyping to Tuberculosis Prevention and Control*. [Online]. Availabel: http://www.cdc.gov/tb/programs/genotyping/manual.htm [Date Last Accessed, 12/01/2014].

[58] R. E. Espinoza, private communication, Oct. 2009.

[59] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *J. of the ACM (JACM)*, vol. 36, no. 4, pp. 929–965, Oct. 1989.