CHAPTER 3

Estimating the Relative Contribution of Transmission to the Prevalence of Drug Resistance in Tuberculosis

Tanzila K. Chowdhury, Laurence A.F. Park, Glenn Stone, Mark M Tanaka, and Andrew Francis

3.1 INTRODUCTION

Tuberculosis (TB) is a respiratory infectious disease caused by the bacterium *Mycobacterium tuberculosis*. The World Health Organization (WHO) estimates that about 1.5 million people died in 2020 due to tuberculosis [2020]. The TB latent and infectious period span long time intervals (years on average) and reduce rapidly two weeks after effective treatment is initiated [Ozcaglar et al., 2012]. The most common first-line drugs used to treat TB are rifampicin, isoniazid, ethambutol, streptomycin, and pyrazinamide, but the acquisition of resistance to these drugs has been a growing problem [Diriba et al., 2013]. Indeed, multi–drug-resistant tuberculosis (MDR TB), defined as resistance to isoniazid (INH) and rifampicin (RIF), is an increasing global problem [Espinal, 2003]. There are even some cases of "extensive" drug resistance (XDR), defined as MDR cases

DOI: 10.1201/9781003292357-3

with additional resistance to fluoroquinolone and at least one second-line injectable agent (kanamycin, amikacin, or capreomycin) [Diriba et al., 2013]. Once tuberculosis becomes extensively resistant, it requires very expensive and extensive treatment [Wright et al., 2006]. Compounding the challenge, less affluent countries bear the highest burden of TB, and people with HIV are more likely to develop active TB. In 2020, the 30 highest TB-burdened countries accounted for 86% of new TB cases. Eight countries account for two-thirds of the total, with India leading the count, followed by China, Indonesia, the Philippines, Pakistan, Nigeria, Bangladesh, and South Africa [WHO, 2020].

Understanding the mechanisms and patterns of resistance evolution is both urgent and essential to controlling TB spread. There are two possible ways for a patient to have acquired drug-resistant tuberculosis. The first is through acquiring resistance while infected (resistance evolution or treatment failure), and the second is by transmission (infection with an already-resistant strain). We are interested to know how the state of being drug sensitive or drug resistant is reached. For example, if we observe five resistant cases in an outbreak sample, these resistant cases either occurred via transmission or treatment failure. It is possible that only one person acquired the resistance because they did not administer their treatment properly (treatment failure) and then transmitted a resistant strain to four other people. This is only one of the five possibilities, and another extreme possibility is that all cases of resistance were acquired independently. If public health authorities know which way of acquiring infection with a resistant strain (transmission or treatment failure) is more common, the public health system can use this to take the necessary steps to stop the spread of resistant TB. That is, if we find that resistance was mostly acquired through treatment failure, the health system needs to ensure that effective treatment procedures are in place. On the other hand, if resistance was acquired mainly due to transmission, there needs to be a more robust system in place to control TB transmission.

The purpose of this chapter is to describe an efficient method to estimate the relative contributions of transmission, and treatment failure, to the spread of drug resistance in an outbreak. Luciani et al. [2009] estimated the relative contribution of transmission to the spread of TB drug resistance using approximate Bayesian computation (ABC) and a transmission-mutation model, which produced posterior distributions of key parameters of interest. More recent studies, such as Rodrigues et al. [2018], have also used ABC to look at the relative contribution problem

in the context of multiple drug resistance. However, the process of ABC is computationally heavy and hence time-consuming, taking days to provide a result and longer to fine-tune the model. In this chapter, we investigate a simple method for approximating the parameters of a TB outbreak that produces results in minutes. We also evaluate the accuracy of the computed parameters under different conditions. The contribution of our research is a simpler model that produces quick estimation and predicts similar results to the Luciani et al. [2009] analysis. No other studies are known that efficiently estimate the relative contribution of transmission and evolution in the spread of TB drug resistance. The two main advantages of the method described in this chapter are its efficiency and simplicity of calculation, relative to a method such as ABC. If a healthcare worker wants to understand the TB transmission mechanism to decide on the steps/treatment to control TB resistance, getting a quick result will be an advantage over the time and expertise required to get results using an ABC approach. Additionally, people with minimal programming knowledge can use the method we describe; it does not need field experts to estimate the relative contributions to TB transmission.

We will use data from several published sources: Bolivia [Monteserin et al., 2013], Tanzania [Kibiki et al., 2007], Cuba [Diaz et al., 1998], and Chad [Diguimbaye et al., 2006]. These sources provide essential tuberculosis genotypic information and drug resistance information for several first-line drugs: isoniazid, rifampicin, pyrazinamide, streptomycin, and ethambutol. Each dataset has several clusters of isolates and information about their phenotype (drug resistance).

The layout of the chapter is as follows. Section 3.2 describes the structure and the visualisation of the data according to their genotype and drug resistance information. We will define a *resistance acquisition graph* for a single cluster and the most probable event history of that cluster in Sections 3.3 and 3.4. Once we have the resistance acquisition graph of a cluster, we can extend this process to a set of clusters, as explained in Section 3.5. This allows us to achieve the chapter's key goal, to make estimates of the proportion of drug resistance cases that have arisen from the transmission. We have included some analysis of our method in Section 3.6.

3.2 SPOLIGOFORESTS WITH DRUG-RESISTANCE INFORMATION

3.2.1 Spoligoforests

While several genotyping methods can be used to characterise variation in bacterial pathogens, this study is based on a technique known as spacer

oligonucleotide typing or spoligotyping. Spoligotyping is a rapid polymerase chain reaction (PCR)-based method for genotyping strains of the TB complex [Kamerbeek et al., 1997], developed to provide information on the structure of the direct repeat (DR) region in individual TB strains and different members of the TB complex [Streicher et al., 2007]. This method determines the presence or absence of 43 different spacers. The binary string representing the pattern of the presence or absence of spacers in a given isolate constitutes a spoligotype. The data used for this research are based on tuberculosis isolates typed with spoligotyping (see Figure 3.2) and phenotypic information relating to antibacterial drug resistance. While more sophisticated genotyping techniques such as whole genome sequencing (WGS) are now available, spoligotyping remains widely used in public health and epidemiological contexts, because it is cheap, fast, and reliable, and in many countries the cost of WGS means it is not widely available.

Given a dataset of spoligotyped isolates, we can form clusters of isolates with identical spoligotypes and then construct *spoligoforests* with directed edges between clusters [Reyes et al., 2008]. Spoligotyping information is used to construct a spoligoforest of the outbreak sample. Inside a spoligoforest, an arrow goes from cluster *A* to cluster *B* if *A* is the parent of *B*; that is, the genotype of cluster *B* evolved from that of cluster *A*. Assuming there is no homoplasy (when a trait has been gained or lost independently over the course of evolution), a spoligotype can only arise from a single parent spoligotype. However, multiple child clusters can appear from a single parent cluster. A sample spoligoforest is shown in Figure 3.1.

Aside from the genotype, which here is defined by spoligotyping, each cluster in the spoligoforest may have both drug-sensitive and drug-resistant

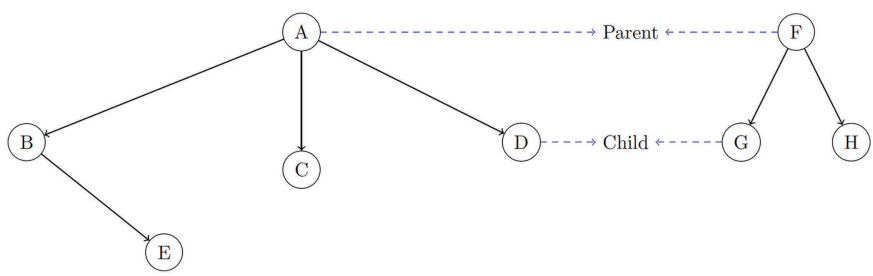
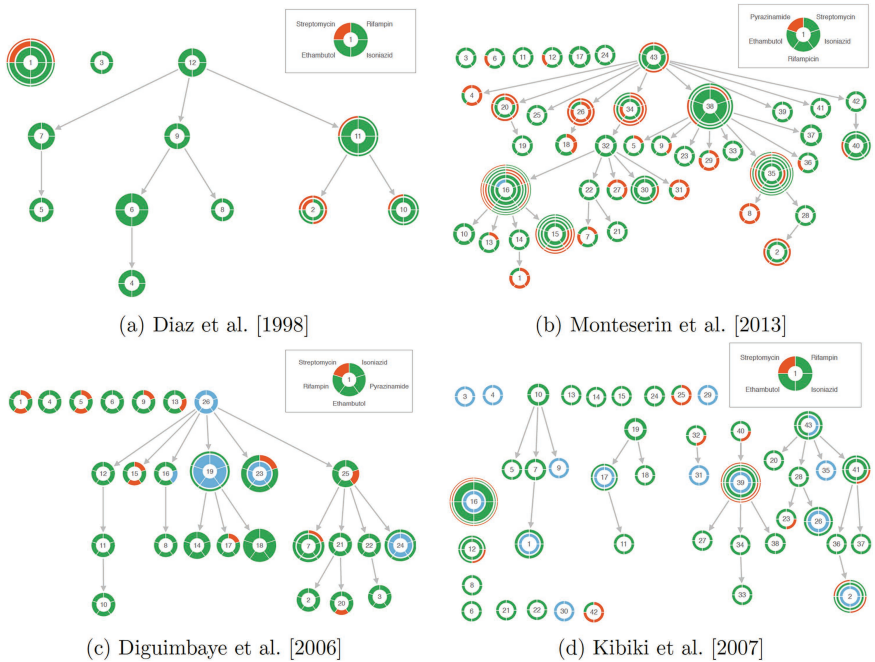


FIGURE 3.1 A sample spoligoforest. Here, A is the parent cluster of B, C, and D clusters, and F is the parent of G and H. B is the parent of child cluster E. Genotype of a child cluster is a genetic evolution from the parent cluster.



(a) Diaz et al. [1998]

(b) Monteserin et al. [2013]

(c) Diguimbaye et al. [2006]

(d) Kibiki et al. [2007]

FIGURE 3.2 Spoligoforests for the datasets taken from Diaz et al. [1998], Monteserin et al. [2013], Diguimbaye et al. [2006], and Kibiki et al. [2007] produced using the MERCAT package [Aandahl et al., 2020]. Each disc represents a cluster of isolates with the same spoligotype, and arrows between clusters represent likely single-step spoligotype mutations. The area of each disc represents the size of the cluster, segments of the discs illustrate drug resistance states for drugs as shown in each legend, and concentric bands indicate isolates with the same resistance profile.

isolates. That is, the cluster is defined by a single genotype but may contain more than one phenotype.

3.2.2 Visualising Spoligoforests with Drug Resistance Information

For a given set of isolates typed with spoligotyping, one can construct a spoligoforest following the methods of Reyes et al. [2008]. This represents a possible mutational history, showing single step directional edges and resolving potential homoplasy to choose a single parent for each cluster (using the method of Ozcaglar et al. [2012]).

While in this study we focus on the simple phenotypes of "sensitive" or "resistant" in relation to a given drug, the methods in MERCAT [Aandahl et al., 2020] allow the visual representation of a range of possible drug resistance "profiles" for each cluster of isolates with identical spoligotype. We have used spoligotyping and phenotypic (drug resistance) information from published datasets. Using this information, MERCAT can construct spoligoforests for the sample. Figure 3.2 shows such spoligoforests of four different datasets labelled accordingly: Diaz et al. [1998] (12 clusters), Monteserin et al. [2013] (43 clusters), Diguimbaye et al. [2006] (26 clusters), and Kibiki et al. [2007] (43 clusters). These graphs were produced using MERCAT [Aandahl et al., 2020]. A summary of all four datasets showing resistance information is given in Table 3.1. These datasets have drug resistance information for five different drugs, rifampicin (RIF), isoniazid (INH), ethambutol (EMB), streptomycin (STR), and pyrazinamide (PZA).

Each node in the spoligoforest produced by MERCAT denotes a genotype cluster. Each concentric band indicates isolates in the cluster with the same drug resistance "profile", and the area of the band is proportional to the number of isolates with that profile (an isolate carrying a drug resistance "profile" means it is resistant to a subset of drugs and sensitive to the remaining drugs). The resistance status of isolates in the cluster with respect to any given drug is then represented in a "pizza slice". The green label indicates sensitivity, the orange label indicates resistance, and the blue label indicates "not identified" cases.

3.3 THE RESISTANCE ACQUISITION GRAPH FOR A SINGLE CLUSTER

This section describes our approach for a single cluster of isolates with identical genotypes. We summarise each cluster with two numbers: drug-sensitive cases (i) and drug-resistant cases (j) as an ordered pair (i, j).

TABLE 3.1 Numbers of Resistant Isolates in the Datasets from Diaz et al. [1998], Monteserin et al. [2013], Diguimbaye et al. [2006], and Kibiki et al. [2007]. We have information for five different drugs collected from these datasets: rifampicin (RIF), isoniazid (INH), ethambutol (EMB), streptomycin (STR), and pyrazinamide (PZA). Pan-sensitive refers to sensitivity to all drugs. Note that some isolates carry resistance to more than one drug.

Drug	Diaz	Monteserin	Diguimbaye	Kibiki
RIF	3	6	0	3
INH	1	8	9	11
EMB	0	7	4	3
STR	16	4	0	4
PZA	—	1	3	—
Pan-sensitive	58	21	19	98
Total isolates	74	35	32	110

For instance, a cluster with one sensitive case and two resistant cases is denoted by (1, 2). Each cluster has several possible histories, beginning with a single source case. We represent these possible historical paths by a *resistance acquisition graph*, which is a directed graph whose vertices represent the state of clusters of one strain. This figure shows all the possible paths by which the cluster may have arisen from a single source case under the assumption that all cases are observed.

We will establish several rules to calculate the "most probable events" (MPEs) on the graph for different clusters. Using this approach, we will determine the most probable evolutionary events of different cluster phenotypes and calculate the frequencies of various evolutionary events (transmission and treatment failure). In the next section, we will combine information from each component cluster in the spoligoforest.

The directed edges in the graph indicate two types of events: the acquisition of drug resistance and the transmission of the pathogen. Edges marking a treatment failure event are labelled A (for acquisition) and indicate the movement of an isolate from sensitive status to resistant status: $(i, j) \rightarrow (i - 1, j + 1)$. Such edges are only possible if the cluster has sensitive cases: if $i \geq 1$. Edges marking a resistance transmission event indicate either transmission of a sensitive case ($(i, j) \rightarrow (i + 1, j)$) or a resistant case ($(i, j) \rightarrow (i, j + 1)$) and are labelled T_s and T_r, respectively. The edges described previously allow only three types of vertices (cluster). These vertex types are: all sensitive $(i, 0)$, all resistant $(0, j)$, or mixed (i, j). The possible edges for each of these types are shown in Figure 3.3.

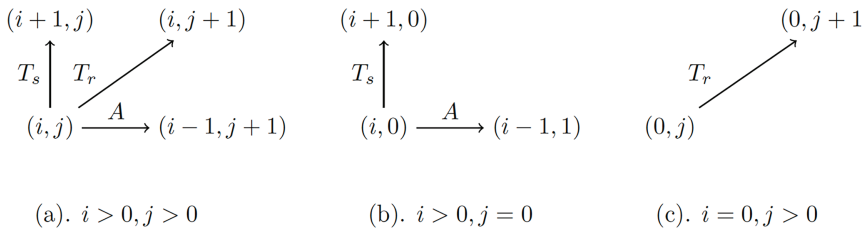


FIGURE 3.3 General rules for all possible events of three different types of clusters. Here, (a) represents a cluster with both sensitive and resistant cases, (b) represents clusters consisting of sensitive cases only, and (c) represents clusters with resistant cases only. All types of events (T_s, T_r, and A) are possible for (a), whereas only T_s and A are possible for (b), and only T_r is possible for (c).

Note that in this model, we assume that the source of a cluster is a single isolate, in accordance with the infinite alleles model (no homoplasy) [Reyes et al., 2012]. We also assume that resistance, once gained, is not lost (following Luciani et al. [2009]). This means that a cluster with sensitive cases *must* have a sensitive case as the source.

As an example, the resistance acquisition graph for the cluster (1, 2) is shown in Figure 3.4 (right). Three possible paths can lead a single sensitive case (1, 0) to a final position (1, 2). Depending on which path we take, the number of transmission and evolution events are different. Two of these paths include two sensitive transmissions (T_s) and two resistance evolution (A) events. Another path includes one sensitive transmission (T_s), one resistant transmission (T_r), and one resistance evolution (A) event. The resistance acquisition graph for a cluster (0, 5) is shown in Figure 3.4.

Each cluster is associated with its own resistance acquisition graph within the spoligoforest. Once we can calculate the relative contribution of resistance transmission and treatment failure in a cluster, we can extend that approach to all possible clusters in the spoligoforest. This way, we will be able to calculate the proportion of resistance due to transmission/treatment failure for the whole sample and then for the entire outbreak by accounting for the weighting factor. Figure 3.5 shows a sample spoligoforest and the approach to studying various events within an outbreak. In Section 3.5, we will discuss how to combine information across clusters in the spoligoforest to study the whole outbreak.

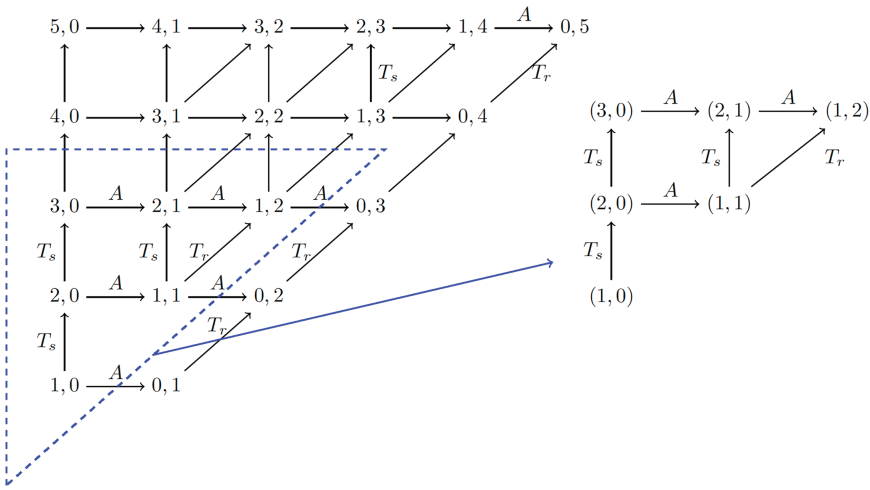


FIGURE 3.4 The resistance acquisition graph for a cluster with five resistant strains (0, 5) is shown on the left. We omit parentheses for simplicity. Note that *all* resistance acquisition graphs for clusters of sizes up to 5 are included in this figure. The graph for a cluster with one sensitive strain and two resistant strains (1, 2) is shown on the right. Here, (1, 0) is the source (bottom left), and (1, 2) is the final position of this cluster (top right corner). Arrows inside the graph indicate the evolution of drug resistance (labelled A) or the transmission of sensitive or resistant strains (T_s and T_r, respectively).

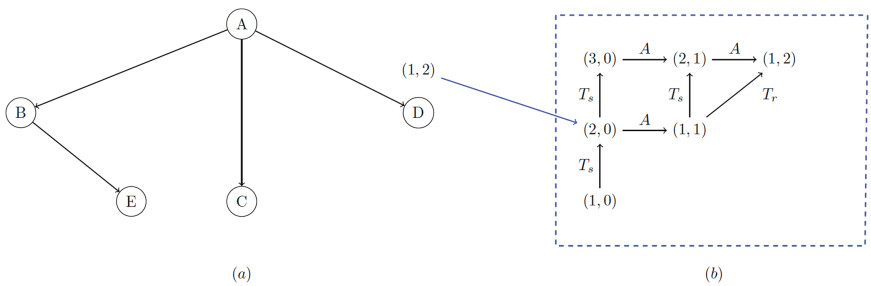


FIGURE 3.5 A sample spoligoforest (a). Within this spoligoforest, we can construct a resistance acquisition graph for each cluster. An example of the resistance acquisition graph of cluster D is shown here (b). This cluster has one sensitive case and two resistant cases.

3.4 HISTORY OF A CLUSTER

In the previous section, we saw that several possible paths through the resistance acquisition graph lead to the final observed cluster. While each path in the graph may correspond to many parallel sequences of events (for instance, the edge (2, 0) → (1, 1) could correspond to the acquisition of drug resistance by either of the two isolates in the initial cluster), in this section, we consider which *individual* path is the most probable.

The probability of each path through the resistance acquisition graph is proportional to the product of the probabilities of the events that occurred on each step of that path. Let α, τ_s, τ_r be the rates of evolution/treatment failure, sensitive transmission, and resistant transmission per individual per unit time (year), respectively. The probability of a particular path to the cluster (i, j) is determined by the frequencies of the different events along the path. For instance, the path (1, 0) → (2, 0) → (1, 1) → (1, 2) in Figure 3.4 has one sensitive transmission, one resistant transmission, and one evolution of drug resistance event, so the probability of the path is given by the product $\tau_s\tau_r 2\alpha$. The number 2 before α is included because this represents the edge probability of (2, 0) → (1, 1), and there are two sensitive cases in that edge, both of which can acquire resistance. However, the path (1, 0) → (2, 0) → (1, 1) → (2, 1) → (1, 2) in Figure 3.4 has two sensitive transmissions and two evolutions of drug resistance, so the probability of the path is given by the product $\tau^2(2 \times 2)\,\alpha^2$. Therefore, we need to consider the frequency of the events in the path as well as, sometimes, the multiplicities of sensitive or resistant cases in each position of the path. If we can compute the probability of every path in the resistance acquisition graph, we can determine the most probable history of an observed cluster.

However, this requires calculating probabilities of each edge, and the number of edges increases rapidly as the cluster sizes increase. As an alternative, we will look at the most probable events along the path, which ignores the frequency of different cases at each position along the path. This method produces results efficiently and provides an easily computable proxy for the most probable path of an observed cluster. In future work, we will compare this proxy with a more complete study of the most probable path through the resistance acquisition graph, but for the present, we will assess this proxy against known parameters in simulations (Section 6).

For the specific example shown in Figure 3.4 (right), there are three paths from (1, 0) to (1, 2), and each path has probability proportional to

either $\tau^2\alpha^2$ or $\tau_s\alpha\tau_r$. We summarise these frequencies in a vector of three integers so that we represent the probability $\tau^a\tau^b\alpha^c$ by the triple (a, b, c) of exponents, where a, b, and c are the frequency of sensitive transmission, resistant transmission, and treatment failure events, respectively. The most probable events for a cluster (i, j) will correspond to a particular triple of exponents, which we denote $\varepsilon_{(i,j)}$. This triple depends on whether the root of the cluster is a sensitive or resistant case ((1, 0) or (0, 1)), so we will often make this clear with a superscript, $\varepsilon^{(1,0)}_{(i,j)}$ or $\varepsilon^{(0,1}_{(i,j)}$.

The main result of this section is the following proposition, which gives the most probable event history to any cluster from a sensitive source case.

Proposition 1. *The most probable events history of the cluster* (i, j) *from source* (1, 0) *with different rates* $\tau_s = 0.66$, $\tau_r = 0.60$, *and* $\alpha = 0.01$ *has event frequencies for* T_s, T_r, *and A events given as follows:*

$$\varepsilon^{(1,0)}_{(i,j)} = \begin{cases} (i, j-1, 1) & \text{if } i \geq 0, j > 0, \text{ and} \\ (i-1, 0, 0) & \text{if } i > 0, j = 0 \end{cases} \tag{1}$$

Proof. In the resistance acquisition graph (see Figure 3.4), the (i, j) cluster is j steps to the right and $i + j - 1$ steps above the source isolate (1, 0). There are two types of steps to move above: sensitive transmissions (T_s) and resistant transmissions (T_r), and two types of steps to move right: resistant transmission (T_r) and resistance evolution (A) events. Suppose that on a given path π, there are (a) sensitive transmissions (T_s), (b) resistant transmissions (T_r), and (c) resistance evolution (A) events. Then this implies $a + b = i + j - 1$ (the number of steps up), and $b + c = j$ (the number of steps to the right). Making b the object, we have

$$a = \cdot i + j - 1 \cdot - b$$

$$c = \cdot j - b$$

Note that b is between 0 and j but is strictly less than j since an initial resistance evolution must occur so that $0 \leq b \leq j - 1$ and hence $c > 0$. We can calculate the value, which is proportional to the probability of this path, $P(\pi)$, as follows.

$$P(\pi) = \alpha^{j-b}\tau_s^{i+j-1-b}\tau_r^b = \alpha^j\tau_s^{i+j-1}\left(\frac{\tau_r}{\alpha\tau_s}\right)^b \tag{2}$$

Since i and j are fixed by the cluster, maximising this probability reduces to choosing a path with the maximal number of resistant transmissions (b), as long as the ratio $\frac{Tr}{\alpha Ts}$ is more than 1.

Previous studies showed that, in general, the rate at which resistant transmissions occur is significantly higher than the rate of resistance acquisition events, or $\tau_r > \alpha$ [Luciani et al., 2009]. That paper estimated the rates of sensitive transmission, resistant transmission, and resistance acquisition to be 0.66, 0.6, and 0.01, which forces $\frac{Tr}{\alpha Ts}$ to be more than 1.

Assuming $j > 0$, the highest possible value of b is $j - 1$, and choosing this value forces $a = i$ and $c = 1$, as required by the statement of the proposition. On the other hand, when $j = 0$, the relation $c = j - b$ forces c to be negative unless $b = 0$. Consequently, $b = c = 0$, and $a = i + j - 1 - b = i - 1$, giving $\varepsilon_{(i,j)}^{(1,0)} = (i - 1, 0, 0)$, as required by the proposition.

In some of what follows, we will consider the situation in which there is a single resistant case as the source for a cluster (such a cluster necessarily only consists of resistant isolates). In that case, there will be one less acquisition of drug resistance than with a sensitive case as a source: we remove the step $(1, 0) \to (0, 1)$. This argument gives the following lemma:

Lemma 4.2. *When the source is a single resistant case* $(0, 1)$ *for a cluster of resistant strains* $(0, j)$*, there will be one less drug resistance acquisition than with the sensitive source. In this case, for* $i = 0$, $j > 0$, $\varepsilon_{(i,j)}^{(1,0)} = (0, \cdot j \cdot - \cdot 1, \cdot 0)$.

3.5 BEYOND A SINGLE CLUSTER

In this section, we will extend our method of calculating the transmission proportion for a single cluster to the whole outbreak. Moving from the single-cluster analysis given in the previous sections to evaluating a multi-cluster outbreak requires summing over the clusters and accounting for connections between clusters in the spoligoforest. This section shows how to do this, beginning with the issues around choosing a source for each cluster. Appointing a source for a cluster is essential, as we need to account for the event that has happened to start a new cluster. If the source of an observed cluster is (1, 0), a sensitive transmission happens from the ancestor cluster to the source of the observed cluster. However, if the source of an observed cluster is the resistant case (0, 1), then a resistance acquisition

event occurred from the ancestor cluster to the source of the observed cluster. This is explained in detail in section 3.5.2.

3.5.1 Assigning the Source Isolate

In studying a single cluster, we accounted for two alternatives for the source isolate: it could be either drug sensitive (1, 0) or drug resistant (0, 1). In what follows, we will describe how this choice of source can be forced by the wider context of the cluster or chosen based on the parsimonious approach used in this research.

We have assumed that resistance cannot be lost for an individual isolate. This has an effect on the source of a cluster in some situations: if the cluster has any drug-sensitive cases, then the source of the cluster must have been sensitive. On the other hand, if the cluster contains only resistant cases, then the assumption that resistance cannot be lost does not rule out either possible source. However, that assumption can still have an impact on the source of the cluster, depending on the descendants of the cluster.

Specifically, if we know that one of the descendants of an all resistant cluster has a sensitive root source, then since the descendant cluster's genotype evolved from the ancestor cluster's genotype, this source case was originally a member of the ancestor cluster. This forces the source of the ancestor cluster to have been a sensitive case. In other words: a sensitive source in the descendant forces a sensitive source in the ancestor.

If none of the descendant clusters of an all resistant cluster has sensitive sources, then we need to look at the ancestor cluster. If the ancestor cluster has no resistant cases, then somewhere in between the clusters, the acquisition of drug resistance occurred. Therefore, for simplicity, we will assume that the source of these types of clusters ((0, j) cluster with (0, j) descendant and (i, 0) ancestor) is a sensitive source.

Finally, suppose our cluster has no sensitive case, none of its descendant clusters has a sensitive source, and the ancestor cluster consists of resistant cases ((0, j) cluster with (0, j) descendant and (i, j) or (0, j) ancestor); then the most probable event path is from a source that is a single resistant case.

To summarise, the rules we follow to assign a source to a cluster (i, j) are described in Proposition 2.

Proposition 2. *Assuming that resistance once attained, cannot be lost and that there is no homoplasy, the source of a cluster* (i, j) *that has arisen from the most probable event path from a single isolate is given in the following cases:*

(*1*) *Cluster* (*i*, *j*) *or* (*i*, 0)*: If there are sensitive cases in a cluster* ($i > 0$)*, the source is a sensitive case* (1, 0).

(*2*) *Cluster* (0, *j*)*: If there is no sensitive case in a cluster* ($i = 0$)*, the source is dependent on the ancestor and descendant clusters, as follows:*

(*a*) *Descendant cluster* (*i*, *j*) *or* (*i*, 0)*: If there is a descendant cluster with sensitive cases* ($i > 0$)*, the source of the cluster* (0, *j*) *is a sensitive case* (1, 0).

(*b*) *Descendant cluster* (0, *j*)*: If the descendant cluster consists of resistant cases only:*

(*i*) *Ancestor cluster* (*i*, 0)*: If the ancestor cluster consists of sensitive cases only, the source of the cluster* (0, *j*) *is a sensitive case* (1, 0).

(*ii*) *Ancestor cluster* (*i*, *j*) *or* (0, *j*)*: If the ancestor cluster consists of resistant cases, the source is a resistant case* (0, 1).

This proposition gives an algorithm for assigning drug resistance status to source isolates of clusters in a spoligoforest. We first determine the source of clusters that have no descendants. Their ancestors can then have their sources determined, and so on. The decision behind the allocation of sources moves from the descendant to the ancestor.

3.5.2 Event Calculations for the Whole Sample

This section describes how to compute the totals of different events (sensitive and resistant transmission and treatment failure) across the whole sample. The next step is summing the totals of each event from each cluster once sources have been assigned according to the previous section. However, we must also account for events that give rise to the source case, which may involve looking at the ancestor of each cluster.

The source case of a cluster is the first with its particular genotype. This evolution of genotype is assumed to occur independently of resistance evolution and transmission dynamics. Therefore, the source case acquired its resistance status from an event, and this is what we need to count.

If the source case is *sensitive*, then it must have arisen from a sensitive transmission T_s. This transmission would have occurred within the ancestor cluster before the genotype evolution but is not counted there: we will count it in the cluster where the source case occurs.

We have denoted the count of events for the most probable event path in the cluster with a sensitive source by the triplet $\varepsilon_{(i,j)}^{(1,0)} = (a, b, c)$, where a, b, and c are the number of sensitive transmission, resistant transmission, and treatment failure events, respectively. We will denote the augmented count by $\varepsilon^{\circ(1,0)}_{(i,j)}$ which includes the event for the source isolate. The previous discussion shows:

$$\varepsilon^{\circ(1,0)}_{(i,j)} = \varepsilon^{(1,0)}_{(i,j)} + (1,0,0)$$

Suppose the source isolate is drug *resistant*. In that case, there are two possible ways it could have arisen from a resistant transmission T_r (from another resistant isolate) or resistance evolution A (from a sensitive isolate). To select which of these to count, we need to inspect the ancestor cluster.

For a cluster to have a resistant source, there are two cases for the ancestor cluster: it could consist entirely of resistant cases (0, j) or of a mixture of both (i, j) (Proposition 1). If the ancestor cluster consists of entirely drug-resistant cases, it must have arisen from a resistant transmission. If the ancestor cluster contains both sensitive and resistant cases, then the source case of the descendant may have arisen from either a drug-resistant transmission or an evolution. However, considering both clusters as a single system, the most probable event path would have a single drug evolution, and this would already be counted in the ancestor cluster. So we take the source isolate to have arisen from a resistant transmission.

In other words, $\varepsilon^{\circ(1,0)}_{(i,j)} = \varepsilon^{(1,0)}_{(i,j)} + (1,0,0)$ if the ancestor cluster is all resistant or mixed.

In summary, we have

$$\varepsilon^{\circ(1,0)}_{(i,j)} = \begin{cases} (i,0,0 & \text{if the cluster is all sensitive and otherwise.} \\ (i+1, j-1,1 & \end{cases}$$

$$\varepsilon^{\circ(1,0)}_{(i,j)} = \varepsilon^{(1,0)}_{(i,j)} + (1,0,0) \quad \text{for all cases}$$

Note that we have assigned a sensitive source for a root cluster of the spoligoforest, which has no ancestor clusters. In this case, there is not enough information about the origin of the source case, and for simplicity, we assume that the source is a sensitive case (1, 0).

The inference described here is deterministic, allowing us to infer the numbers of events directly from cluster structure. In particular, the

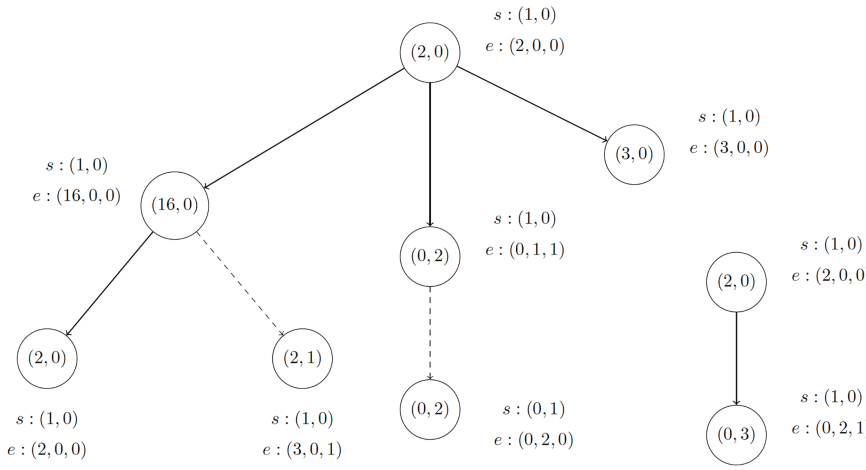


FIGURE 3.6 Spoligoforest for data from Monteserin et al. [2013], focusing on resistance to isoniazid. This graph corresponds to the second row of the Monteserin column in Table 3.1. Each circle represents a cluster with (*i, j*), meaning *i* sensitive and *j* resistant cases. The source of each cluster (*s*: (1, 0)/(0, 1)) and the frequency of events ($e = (T_s, T_r, A)$) are shown next to each cluster.

numbers of resistant and sensitive cases in an isolated cluster determine the proportion of resistant cases due to transmission. In the generic case of a cluster with *i* sensitive and *j* resistant cases ($i, j > 0$), the proportion of resistant cases due to transmission is inferred to be $\frac{b}{b+c} = \frac{j-1}{j}$ (Proposition 1).

An example of assigning sources to clusters, as described in Proposition 2, is shown in Figure 3.6. This figure accommodates all the rules we have introduced to calculate different events for the whole spoligoforest.

3.5.3 Accounting for Sampling

The data we observe represent only a sample of an underlying infected outbreak. To estimate the parameters of an infected outbreak, the next step is to account for sampling weights so that our sample represents the infected outbreak. We assume that the process of sampling follows a binomial distribution with two outcomes: success indicates we have successfully sampled the case from an outbreak (we observe this in the sample), and failure indicates that we have failed to sample the case from an outbreak. We assume that sampling is not biased with respect to resistance and that there is an equal chance of sampling a sensitive or resistant case.

Suppose the true infected population size is n, with i sensitive and j resistant cases, and the probability of success (being sampled) is λ. Feldman and Fox [1968] show that for k ($k = 1, \ldots, r$) independent binomial random variables (X_k) with $B(n, \lambda)$, the maximum likelihood estimate for the true infected population n satisfies:

$$\prod_{k=1}^{r}\left(1-\frac{x_k}{\hat{n}}\right)=(1-\lambda)^r \tag{3}$$

We use $\hat{n}$ as the estimation of the true infected population n. For our distribution, $r = 1$, as we deal with just one independent random variable. Therefore:

$$\begin{aligned} 1-\frac{x}{\hat{n}} &= 1-\lambda \\ \hat{n} &= \frac{x}{\lambda} \end{aligned} \tag{4}$$

Therefore, we can scale up the observed cluster (i, j) by dividing the sampling proportion. That is, for sampling proportion $\lambda \in (0, 1]$, we scale up each cluster so that (i, j) becomes $\left(\frac{i}{\lambda}, \frac{j}{\lambda}\right)$ In particular, if a cluster has zero sensitive or resistant cases, this stays zero. As we are dealing with TB cases, we have rounded the result for simplicity.

3.5.4 The Effect of Sampling on the Proportion of Resistance due to Transmission

To calculate the proportion of resistance due to transmission for an outbreak, we need to consider the sampling behind the data. Keeping the proportions of sensitive and resistant cases constant in a cluster, as the cluster size grows, the inferred proportion of resistance due to transmission also grows.

Suppose the true cluster size is m, with j resistant cases. If the sampling proportion is λ, then the cluster size of the sample is λm with λj resistant cases. The inferred proportion of resistant cases arising from transmission will be $\frac{\lambda j-1}{\lambda j}=1-\frac{1}{\lambda j}$, while the true proportion should be $\frac{j-1}{j}=1-\frac{1}{j}$ Since $\lambda j \cdot <= \cdot j$, the effect of sampling is to underestimate the proportion of resistant cases due to transmission.

However, we can make a simple correction if we know the sampling proportion or estimate it from the TB yearly reports prepared by WHO. If we have sampled j^t and the estimated sampling proportion is λ, then our estimate of the infected population level proportion of drug resistance due to transmission is $\frac{\left(j^{'}/\lambda\right)-1}{j^{'}/\lambda}=\frac{j^{'}-\lambda}{j^{'}}$. That is, instead of a proportion of $\frac{j^{'}-1}{j^{'}}$, we use $\frac{j^{'}-\lambda}{j^{'}}$. which is just $1-\frac{1}{j}$ since $j^t = \lambda j$.

3.6 RESULTS

3.6.1 Results, Comparison with Simulation

Once we have the MPE model ready, the next logical step is to test the model to check whether it predicts an accurate measure. In this section, we show the results of the method using a simulation study. To do this, we will compare the result of our model with simulated outbreaks. We will simulate hundreds of outbreaks to predict the parameters of interest (resistance due to transmission and treatment failure). In the simulation process, we know exactly how many different types of events (T_r or A) occur throughout the process. We will compare these numbers with the result from the MPE model.

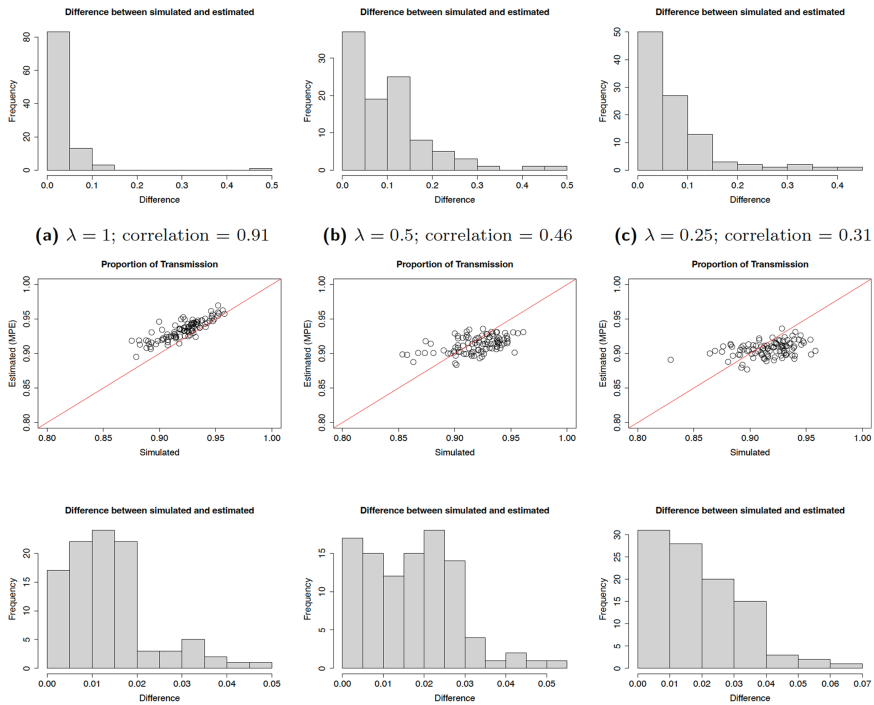
We initialise the simulation from a single sensitive case (1, 0). For any position (i, j) in the simulation, five possible edges indicate five possible events: sensitive transmission, resistant transmission, resistance evolution (treatment failure), sensitive death or recovery, and resistant death or recovery. Using the parameter values estimated by Luciani et al. [2009], we used 0.66, 0.60, 0.01, 0.52, and 0.202 as rates for these five events, respectively. The first three events are conceptually the same as our most probable event calculation method. The last two events are the cure/recovery or deaths for sensitive and resistant cases. These two edges go backwards. A sensitive recovery/death event indicates an isolate's movement from a sensitive status to a cured state, which we do not observe in the graph. In this case, the isolate transforms to a position with one less sensitive case than the initial count: $(i, j) \rightarrow (i - 1, j)$. The same concept applies to a resistant recovery/death event, with one less resistant case after the event: $(i, j) \rightarrow (i, j - 1)$.

We simulate an outbreak from a single sensitive case (1, 0) using the five different types of events mentioned previously. Although each cluster is simulated individually, we have simulated outbreaks consisting of a set of clusters whose sizes are determined using the infinite allele

model (IAM), as described in Luciani et al. [2008]. Given an outbreak size and an estimate of the IAM parameter θ (related to mutation rate and effective population size), the IAM predicts a distribution of cluster sizes. Using the estimation of θ from Luciani et al. [2008], we have used $\theta = 40$ in our calculation process to simulate an outbreak. In the simulation process, we know exactly how many cases are present in each cluster, how many clusters are in the outbreak, and how many resistance transmission and treatment failure events occurred during this process.

The outbreak sizes of the four published datasets we have used in this research range from approximately 800 to 3500. Guided by these, we have simulated different sizes of outbreaks (300–3000). However, for simplicity, only results from outbreak sizes of 300 and 1500 are shown in Figure 3.7. Once we count exactly how many events of each type occur during the simulation, the next step is taking a sample from the simulated outbreak. We use different sampling proportions, with a sampling proportion of 1 (complete sampling), 0.5, and 0.25, shown in Figure 3.7. To account for sampling, we scale up each sample using the sampling factor. Then, we can calculate events using our proposed most probable event calculation method for this scaled sample. As we don't have any genotypic information for this process, for simplicity, we assume that the source of each cluster is a single sensitive case (1, 0).

Once we have the number of each event, we can calculate the proportion of resistant events arising through transmission using both methods (simulation and proposed). Figure 3.7 shows these proportions for two different methods plotted against each other. These figures were produced for an infected outbreak size of 300 (a to c) and 1500 (d to f). We have also simulated other outbreak sizes, such as an outbreak size of 3000. The correlation between the simulated and estimated transmission proportion from an outbreak size of 3000 is very similar to the results from an outbreak size of 1500. Therefore, we have omitted the results from outbreak size 3000. These outbreaks shown in the graph include sampled and non-sampled infections. We have compared these simulated outbreaks with our estimated proportion 100 times for each outbreak size. These graphs show that when the sampling proportion is not small (approximately bigger than 0.3), the proposed method and simulation result shows a linear relationship, and they are very close to each other.



(a) $\lambda = 1$; correlation $= 0.91$ **(b)** $\lambda = 0.5$; correlation $= 0.46$ **(c)** $\lambda = 0.25$; correlation $= 0.31$

(d) $\lambda = 1$; correlation $= 0.85$ **(e)** $\lambda = 0.5$; correlation $= 0.48$ **(f)** $\lambda = 0.25$; correlation $= 0.36$

FIGURE 3.7 Comparison of the proportion of resistance due to transmission (first and third row) and the difference (second and fourth row) between simulated and estimated method is shown in the sets A to F. Three different sampling proportions (λ = 1, 0.5, 0.25) are used for each parameter (transmission proportion and difference). These graphs were produced for outbreak sizes of 300 (a to c) and 1500 (d to f—zoomed graph).

The larger outbreak size (outbreak size 1500, Figure 3.7, d to f) shows a smaller difference between simulated and estimated proportions for all different sampling proportions. However, for outbreak sizes 1500 and higher, both simulated and estimated transmission proportions usually stay between 0.8 and 1.0. Therefore, obtaining an estimated proportion within 0.05 of the simulated proportion doesn't necessarily indicate the accuracy of the method. When we look at the correlation between simulated and estimated transmission proportion in these graphs, we can see that the correlation is highest when the sampling proportion (λ) is 1

(complete sampling) and decreases with sampling proportion (correlation of Figure 3.7(a) and (d) is higher than Figure 3.7(b) and (e); Figure 3.7(b) and (e) is higher than Figure 3.7(c) and (f)). This indicates that the correlation between the simulated and estimated transmission proportion increases as the sampling proportion increases. However, the size of an outbreak doesn't have much impact on the correlation parameter. The only case when it can be an issue is when the outbreak size is small; also, the sampling proportion is small (approximately ≤ 300, Figure 3.7(c)) and is less than 0.3. The estimated transmission proportion still shows a linear relationship with the simulated proportion (correlation coefficient is approximately 0.3); however, some of the values are overestimated in the proposed MPE method. This comparison of the simulation and the proposed method shows that the MPE method is a reasonable estimate of the number of different events when the sampling proportion is not too low.

The weighting process can have an impact on estimating MPE parameters. To understand this, we will discuss an example where we have ten resistant cases in one of the clusters in the sample, and we use 10% sampling (sampling proportion $\lambda = 0.1$); our method assumes that there are 100 resistant cases in the outbreak when we account for scale-up correction. The proportion of resistance arising through transmission is 0.99 for this cluster. However, if we assume the sampling proportion is 50%, the proportion of resistance due to transmission comes down to 95%. Therefore, in the sampling process, we lose some information, and as the sampling proportion decreases, the proportion of resistance due to transmission increases. However, bigger outbreak sizes produce an exceedingly high proportion of resistance due to transmission (close to 1); scaling up does not affect these cases much.

As the outbreak sizes increase, both simulated and estimated transmission proportion increases. This is much more apparent for bigger outbreak sizes, as only a few treatment failure events (one for each cluster) are counted in the whole outbreak using the MPE method. The MPE estimate of transmission proportion $((j - 1)/j)$ clearly approaches 1 with increasing outbreak sizes. This high proportion of resistance due to transmission also reflects the simulation and real outbreaks as the parameter for transmission rate is 0.6 vs the treatment failure rate is 0.01. These parameters were estimated from real data [Tanaka et al., 2006]. As the rate of transmission is much higher than the rate of evolution (treatment failure), as the outbreak grows, the proportion will get close to 1.

Figure 3.8 shows the heat map with contour lines of the correlation and difference between the simulated and estimated transmission proportions. Here, the black contour line follows when there is an equal correlation (Figure 3.8(A)) or difference (Figure 8(B)) between the simulated and estimated transmission proportion. That is, the parameter (correlation coefficient or difference between simulated and estimated proportion) stays constant for each contour line. The contour lines from Figure 3.8(A) indicate that as the outbreak size increases, the correlation between simulated and estimated proportion due to transmission slightly decreases for sampling proportions lower than 0.6. However, there is no significant change in correlation at each level of the graph. From this graph, we can also see that the sampling proportion and the correlation coefficient (between the simulated and estimated transmission proportion) have a positive relationship. As the sampling proportion increases, the correlation between the simulated and estimated proportion (due to transmission) also increases (goes from light yellow to dark red). A high correlation coefficient indicates a good MPE transmission prediction compared to simulated outbreaks. The correlation contour plot shows that the correlation between simulated and estimated proportions is more than 0.5 if the sampling proportion is above 50%. As the colour darkens vertically, not horizontally, this indicates that the correlation coefficient is mostly dependent on the sampling proportion of the outbreak.

Figure 3.8(B) shows that the absolute difference between the simulated and estimated transmission proportion decreases as the sampling proportion increases. Here, the black contour line follows when there is an equal difference between the simulated and estimated transmission proportion. From this graph, we can see that the difference decreases as both the sampling proportion and outbreak size increase (colour lightens vertically and horizontally). Generally, the smaller the difference between the estimated and simulated transmission proportions is, the better the MPE estimate is. The absolute difference between the simulated and estimated proportion is less than 5% as long as the outbreak size is bigger than 400. Generally, it is reasonable to assume that the outbreak sizes are bigger than the size 400. This difference graph (Figure 3.8(B)) indicates that the estimated proportions get more accurate as the outbreak sizes increase and the sampling proportion increases. In summary, Figure 3.8 shows that the MPE predicts a reasonable estimate of the resistance proportion due to transmission.

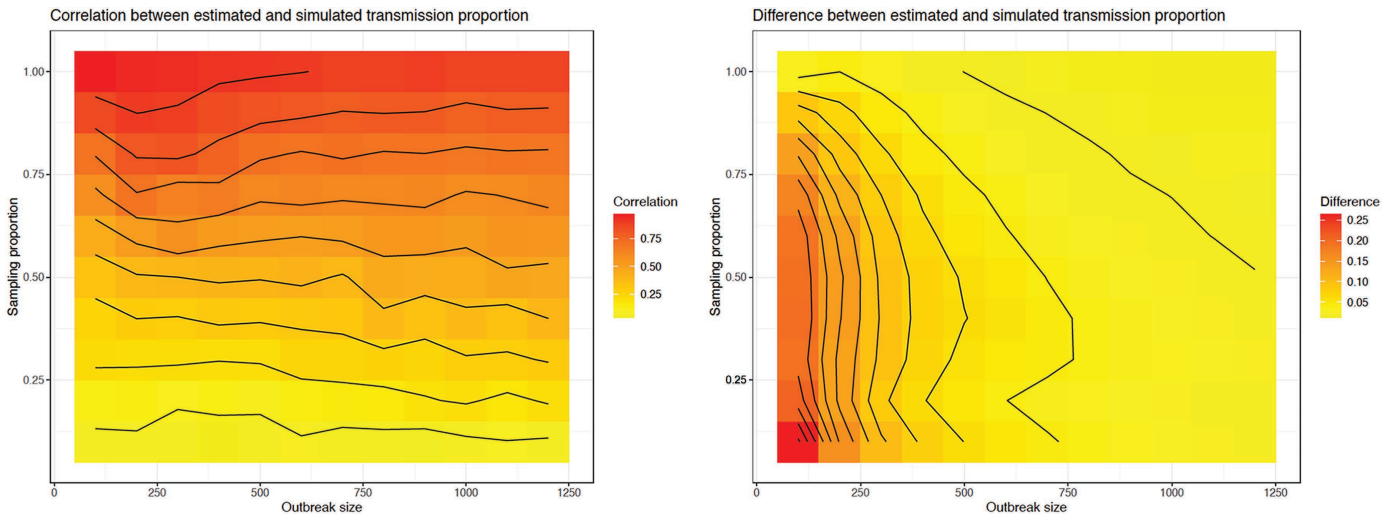


FIGURE 3.8 Contour map comparison of the correlation (A) and difference (B) between the proportion of resistance arising through transmission using the MPE method and simulation. These graphs are produced using the average results of 3000 simulations.

3.6.2 Application to Four Datasets

We are now in a position to apply the method to published datasets to count the events in the most probable event paths leading to the observed data. These data contain resistance phenotype information for several front-line drugs, and we treat these independently so that for each drug, we classify an isolate as "sensitive" or "resistant". Therefore, these datasets contain information on drug resistance status (sensitive/resistant) for each case (people) and their genotype. Using this information, we first develop a spoligoforest. Within the spoligoforest, we assign sources for each cluster and then calculate the number of events occurring in the most probable event path of each cluster in the spoligoforest. This way, we calculate the different number of events and the proportion of resistance due to transmission/treatment failure.

We focus on the *source of resistance cases*, events T_r and A. Resistance transmission event counts for different drugs using the dataset from Monteserin et al. [2013] are shown in the second column of Table 3.2 under the assumption of complete sampling (assuming the sample represents the whole outbreak). This table shows the number of transmission events (T_r) and the proportion of resistance due to transmission (p_t). For instance, in the second column, the RIF row indicates four resistance events occur through transmission for the rifampicin drug, and the proportion of resistance due to transmission is almost 67%.

The next step is to introduce sampling proportion in this section. We calculate the proportion of resistance due to transmission (P_t) for different sampling proportions (λ = 1, 0.5, 0.1, 0.01) for all five drugs, which is shown in Table 3.2. λ is the sampling proportion, where $\lambda = 1$ indicates that the sampling proportion is 100% and we have the whole outbreak dataset sampled (complete sampling). Elaborate results of 10% sampling for each of the four different datasets are shown in Table 3.3.

Table 3.3 shows the proportion of drug resistance due to transmission (T_r) or treatment failure (A) for each drug in each dataset, assuming the sampling proportion is 10%. For instance, we can see that 93% of drug resistance to rifampicin was caused by transmission, which means 7% of resistance was caused by evolution in Diaz et al. [1998] data. The combined result shows that, on average, almost 94% of the resistance was caused by transmission (the last column in Table 3.2). Luciani et al. [2009] estimate the median of the proportion of resistance due to treatment failure to be 0.0270 [0.000, 0.238] for Cuba data and 0.019 [0.000, 0.157] for Estonia data.

TABLE 3.2 Proportion of Resistant Transmission Events Denoted by p_t for Different Sampling Proportions ($\lambda = 1, 0.5, 0.1, 0.01$). Resistant transmission events are denoted by T_r showed for five different drugs: rifampicin (RIF), isoniazid (INH), ethambutol (EMB), streptomycin (STR), and pyrazinamide (PZA). Datasets used for these calculations are collected from Monteserin et al.'s [2013] paper.

λ	1		0.5		0.1		0.01	
Drug	T_r	p_t	T_r	p_t	T_r	p_t	T_r	p_t
RIF	4	0.667	10	0.833	58	0.967	598	0.997
INH	5	0.625	13	0.813	77	0.963	797	0.996
EMB	3	0.429	10	0.714	66	0.943	696	0.994
STR	2	0.500	6	0.750	38	0.950	398	0.995
PZA	0	0	1	0.500	9	0.900	99	0.990

TABLE 3.3 Proportion of Resistant Transmission Events (P_t), Assuming That Collected Datasets Are 10% of the Whole Population. Resistant transmission events are denoted by T_r, shown for five different drugs.

	Diaz		Monteserin		Diguimbaye		Kibiki		Combined	
Drug	T_r	p_t	T_r	p_t	T_r	p_t	T_r	p_t	T_r	p_t
RIF	28	0.93	58	0.97	0	0	27	0.90	113	0.94
INH	9	0.90	77	0.96	83	0.92	101	0.92	270	0.93
EMB	0	0	66	0.94	36	0.90	27	0.90	129	0.92
STR	156	0.98	38	0.95	0	0	37	0.93	231	0.96
PZA	—	—	9	0.90	28	0.93	—	—	37	0.93

That paper estimates that > 90% of resistance cases in Estonia and Cuba are attributable to transmission, indicating that the results in Tables 3.1 and 3.2 are consistent with the findings in Luciani et al. [2009].

3.7 DISCUSSION

Our model finds the most probable events on a path and calculates different numbers of events according to the most probable event path. It shows that there is compulsorily only one evolution event in the most probable event path for each cluster, and the rest are resistant transmission events due to such a low evolution rate. We have introduced an elementary combinatorial method to estimate the proportion of drug resistance due to treatment failure. This is a significantly more efficient method than others, such as ABC [Luciani et al., 2009], and it recovers similar estimates. Compared with the simulation, the MPE estimates are

very close to the actual proportions. Only smaller outbreaks with a small sampling proportion overestimate this parameter. The possible reason for overestimation is explained in section 3.6.1. However, it is reasonable to assume that the outbreak sizes are not very small when an outbreak occurs in an area.

Table 3.3 shows the result for the proportion of transmission and treatment failure events assuming a 10% sampling proportion. This table gives a clear understanding of the percentage-wise result for resistance due to transmission (p_t). Our model estimates that more than 90% of the resistance arises through transmission for all four different datasets for each drug separately, and on average, 94% of the resistance is due to transmission for all the drugs when we combine four datasets results, assuming the sampling proportion is 10%. Estimating an overall standard sampling proportion can be done by looking at estimates of overall disease load from the WHO for a region, compared to the sample sizes in the studies, as done in Tanaka et al. [2006]. However, this study is focused on estimating the relative contribution of TB transmission to the spread of resistance for different sampling proportions and does not claim that a particular sampling proportion is more common. Our result agrees with the conclusions of other studies that tuberculosis drug resistance reported in these studies arises mainly through transmission.

This chapter provides a way to estimate the proportion of drug-resistant cases that arise through transmission or treatment failure in a computationally simple way. In contrast to other methods, notably, the approximate Bayesian computation method takes days to compute. This approach does not require simulation but can be calculated from standard genotype data together with information about the resistance of individual isolates. Methods such as these should help public health bodies to make inferences about their data in a timely manner.

Increasingly, major decisions about the diagnosis and treatment of diseases are taken with the help of data-driven science. This chapter provides an efficient computational method with graph visualisation to estimate the relative contribution of transmission to the spread of TB drug resistance. This model will help understand and improve the actionable steps required to support TB disease control in real environments.

The logical next step is to examine multiple drug resistance. So far, we have looked at only single drug resistance for any isolate for simplicity. Therefore, if an isolate is resistant to any drug, we mark that isolate with

a drug-resistant phenotype. However, we plan to investigate multiple drug resistance and their relationship to each other in the future. The MPE method loses some information throughout the process; however, it is a quick and practical solution. Therefore, we also plan to calculate the expected number of different events in a cluster/outbreak, considering all of the possible paths toward the final cluster. This approach will use more information about the underlying process by integrating all paths.

ACKNOWLEDGEMENTS

The authors thank Arthur Street, Natalia Vaudagnotto, Sangeeta Bhatia, and Zach Aandahl for helping develop and implement the representation of drug resistance profiles used in Figure 3.2 while their study [Aandahl et al., 2020] was in development.

REFERENCES

R Zach Aandahl, Sangeeta Bhatia, Natalia Vaudagnotto, Arthur G Street, Andrew R Francis, and Mark M Tanaka. MERCAT: Visualising molecular epidemiology data combining genetic markers and drug resistance profiles. *Infection, Genetics and Evolution*, 77:104043, 2020.

R Diaz, K Kremer, PEW De Haas, RI Gomez, A Marrero, JA Valdivia, JDA Van Embden, and D Van Soolingen. Molecular epidemiology of tuberculosis in Cuba outside of Havana, July 1994–June 1995: Utility of spoligotyping versus IS*6110* restriction fragment length polymorphism. *The International Journal of Tuberculosis and Lung Disease*, 2(9):743–750, 1998.

Colette Diguimbaye, Markus Hilty, Richard Ngandolo, Hassane H Mahamat, Gaby E Pfyffer, Franca Baggi, Marcel Tanner, Esther Schelling, and Jakob Zinsstag. Molecular characterization and drug resistance testing of *Mycobacterium tuberculosis* isolates from Chad. *Journal of Clinical Microbiology*, 44(4):1575–1577, 2006.

B Diriba, T Berkessa, G Mamo, Y Tedla, and G Ameni. Spoligotyping of multidrug resistant *Mycobacterium tuberculosis* isolates in Ethiopia. *The International Journal of Tuberculosis and Lung Disease*, 17(2):246–250, 2013.

Marcos A Espinal. The global situation of MDR-TB. *Tuberculosis*, 83(1–3):44–51, 2003.

Dorian Feldman and Martin Fox. Estimation of the parameter *n* in the binomial distribution. *Journal of the American Statistical Association*, 63(321):150–158, 1968.

Judith Kamerbeek, Leo Schouls, Arend Kolk, Miranda Van Agterveld, Dick Van Soolingen, Sjoukje Kuijper, Annelies Bunschoten, Henri Molhuizen, Rory Shaw, Madhu Goyal, et al. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *Journal of Clinical Microbiology*, 35(4):907–914, 1997.

Gibson S Kibiki, Bert Mulder, Wil MV Dolmans, Jessica L de Beer, Martin Boeree, Noel Sam, Dick van Soolingen, Christophe Sola, and Adri GM van der Zanden. *M. tuberculosis* genotypic diversity and drug susceptibility pattern in HIV-infected and non-HIV-infected patients in northern Tanzania. *BMC Microbiology*, 1(51), 2007.

Fabio Luciani, Andrew R Francis, and Mark M Tanaka. Interpreting genotype cluster sizes of *Mycobacterium tuberculosis* isolates typed with *IS6110* and spoligotyping. *Infection, Genetics and Evolution*, 8(2):182–190, 2008.

Fabio Luciani, Scott A Sisson, Honglin Jiang, Andrew R Francis, and Mark M Tanaka. The epidemiological fitness cost of drug resistance in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences*, 106(34):14711–14715, 2009.

Johana Monteserin, Mirtha Camacho, Lucía Barrera, Juan Carlos Palomino, Viviana Ritacco, and Anandi Martin. Genotypes of *Mycobacterium tuberculosis* in patients at risk of drug resistance in Bolivia. *Infection, Genetics and Evolution*, 17:195–201, 2013.

Cagri Ozcaglar, Amina Shabbeer, Scott L Vandenberg, Bülent Yener, and Kristin P Bennett. Epidemiological models of *Mycobacterium tuberculosis* complex infections. *Mathematical Biosciences*, 236(2):77–96, 2012.

Josephine F Reyes, Andrew R Francis, and Mark M Tanaka. Models of deletion for visualizing bacterial variation: An application to tuberculosis spoligotypes. *BMC Bioinformatics*, 9(1):496, 2008.

Josephine F Reyes, Carmen HS Chan, and Mark M Tanaka. Impact of homoplasy on variable numbers of tandem repeats and spoligotypes in *Mycobacterium tuberculosis*. *Infection, Genetics and Evolution*, 12(4):811–818, 2012.

Guilherme S Rodrigues, Andrew R Francis, Scott A Sisson, and Mark M Tanaka. Inferences on the acquisition of multi-drug resistance in mycobacterium tuberculosis using molecular epidemiological data. In *Handbook of Approximate Bayesian Computation*, pp. 481–511. Chapman and Hall/CRC, 2018.

EM Streicher, TC Victor, G Van Der Spuy, C Sola, N Rastogi, PD Van Helden, and RM Warren. Spoligotype signatures in the *Mycobacterium tuberculosis* complex. *Journal of Clinical Microbiology*, 45(1):237–240, 2007.

Mark M Tanaka, Andrew R Francis, Fabio Luciani, and SA Sisson. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics*, 173(3):1511–1520, 2006.

WHO. Who 2020. www.who.int/news-room/fact-sheets/detail/tuberculosis, 2020.

A Wright, G Bai, L Barrera, F Boulahbal, N Martin-Casabona, C Gilpin, F Drobniewski, M Havelkov´a, R Lepe, R Lumb, et al. Emergence of mycobacterium tuberculosis with extensive resistance to second-line drugs—Worldwide, 2000–2004. *Morbidity and Mortality Weekly Report*, 55(11):301–306, 2006.