Estimating the Relative Probability of Direct Transmission between Infectious Disease Patients

Sarah E. Van Ness[1, *], Robyn Lee[2], Paola Sebastiani[1], Charles R. Horsburgh[3], Helen E. Jenkins[1, †], and Laura F. White[1, †]

[1]Boston University Department of Biostatistics, Boston, MA
[2]Harvard T.H. Chan School of Public Health, Boston, MA
[3]Boston University School of Public Health Department of Epidemiology, Boston, MA
[*]Corresponding author: Sarah Van Ness, sv1205@bu.edu
[†]These authors contributed equally

Abstract

Estimating infectious disease parameters such as the serial interval (time between symptom onset in primary and secondary cases) and reproductive number (average number of secondary cases produced by a primary case) are important to understand infectious disease dynamics. Many estimation methods require linking cases by direct transmission, a difficult task for most diseases. Using a subset of cases with detailed genetic or contact investigation data to develop a training set of probable transmission events, we build a model to estimate the relative transmission probability for all case pairs from demographic, spatial and clinical data. Our method is based on naive Bayes, a machine learning classification algorithm which uses the observed frequencies in the training dataset to estimate the probability that a pair is linked given a set of covariates. In simulations we find that the probabilities estimated using genetic distance between cases to define training transmission events are able to distinguish between truly linked and non-linked pairs with high accuracy (area under the receiver operating curve value of 95%). Additionally only a small subset of the cases, 10% - 40% depending on sample size, need to have detailed genetic data for our method to perform well. We show how these probabilities can be used to estimate the average effective reproductive number and apply our method to a tuberculosis outbreak in Hamburg, Germany. Our method is a novel way to infer transmission dynamics in any dataset when only a subset of cases have rich contact investigation and/or genetic data.

Introduction

Infectious disease parameters such as the serial interval, defined as the time between symptom onset from primary to secondary case, and the reproductive number, defined as the average number of secondary cases produced by a primary case over the course of their infection, are instrumental in managing outbreaks [1]. These parameters are used to determine how fast an infection is spreading, how well it is controlled, and if public health interventions are working.

For some diseases such as influenza A (H1N1) [1, 2], Severe Acute Respiratory Syndrome (SARS) [3, 4], and Ebola [5, 6], in which disease progression happens shortly after infection, these parameters have been studied extensively. For other diseases such as tuberculosis (TB), the estimates for the serial interval and reproductive number are few and inconsistent [7–9]. Methods to estimate both the serial interval and reproductive number rely on determining which cases are linked by direct transmission. In recent years, pathogen whole genome sequence (WGS) data has been established as a powerful tool to link cases for many diseases [10–22].

Didelot et al. [23] developed a method that infers transmission networks, estimates transmission probabilities, and estimates the serial interval and reproductive number. This method requires a phylogenetic tree as input, which is derived from pathogen WGS data from the observed cases and performs well at inferring transmission events even when when a proportion of cases were unsampled. However, it has been shown that for some diseases, pathogen WGS data is less informative of transmission links for various reasons such as low mutation rates, short time intervals between infections, and within host diversity [24, 25].

Another way to effectively link cases by direct transmission is contact investigations. Contact investigations are often a component of the public health response to an infectious disease outbreak [24, 24, 26–31]. Campbell et al. [24] developed a method to that successfully infers transmission chains using contact investigation data alone and combined with WGS data. However, even contact investigations cannot perfectly determine linked cases as some diseases can be transmitted through casual contact in public locations that would not be captured in typical contact investigations [32, 33]. Additionally, the effectiveness of contact investigations can depend upon characteristics of the disease such as the length of the latency period, the clustering of infections [29], and the willingness of cases to share information about their contacts [31, 34].

The methods outlined by Didelot et al. [23] and Campbell et al. [24] are powerful tools for understanding transmission dynamics using the most discriminating pieces of information currently available: pathogen WGS and contact investigation. However, even these data sources cannot perfectly capture outbreak dynamics, due to the complexity of pathogen genetics and human interactions. Furthermore, WGS data is expensive to obtain and requires bioinformatics expertise and contact investigations are time consuming and require significant human resources. Therefore, these data sources are not universally available for many established infectious disease study samples. We propose a method that uses whatever genetic and/or contact investigation information is available, even if it only available for a small subset of cases, to estimate the probability of direct transmission for all case pairs.

Since it is essentially impossible to know for sure the true infector of a given case, the probability that two cases are linked could be a more informative quantity. We developed a method to predict the relative probability of direct transmission between infectious disease patients using pathogen WGS data and/or contact investigations when these data are only available on a small proportion of cases, paired with demographic, spatial, clinical, and social risk factor data. These probabilities can be used to understand the transmission dynamics of an outbreak and estimate the reproductive number in the absence of a reliable serial interval estimate. We then applied our method to a TB outbreak in Hamburg, Germany that had extensive clinical and epidemiological data.

Methods

Data Structure

Our method starts with a dataset of individuals with the infectious disease of interest for whom various covariate values such as geographic location, clinical information, age, and time of sampling have been observed. Furthermore, additional information such as detailed contact investigation data and/or whole genome sequencing of the pathogen genome is available for at least some of the cases. This subset of individuals serves as the training set to generate the model.

We then transform this dataset of individuals into a dataset of ordered case pairs $(i,\ j)$, where case i was observed before case j. We convert the individual-level covariates $(X_1, X_2, \dots X_p)$ into pair-level covariates $(Z_1, Z_2, \dots Z_p)$ by computing the "distances" between the covariate values for the two cases. These "distances" identify how well the individuals match on the various covariates.

These distances can be dichotomous where a value of 1 indicates that the covariate values match and 0 indicates that they do not match as in Equation 1 for covariate Z_k (where $k \in 1, 2, \dots, p$):

$$Z_{kij} = d(X_{ki}, X_{kj}) = \begin{cases} 1 & \text{if } X_{ki} = X_{kj} \\ 0 & \text{if } X_{ki} \neq X_{kj}. \end{cases} \tag{1}$$

Or the distances could be categorical where different combinations of individual-level values result in r different pair-level values such as:

$$Z_{kij} = d(X_{ki}, X_{kj}) = \begin{cases} 1 & \text{if } X_{ki} = X_{kj} = v_1 \\ 2 & \text{if } X_{ki} = X_{kj} = v_2 \\ 3 & \text{if } X_{ki} = v_1, X_{kj} = v_2 \\ \dots & \\ r & \text{if } X_{ki} = X_{kj} = v_w \end{cases} \tag{2}$$

where $v_1, v_2, \dots, v_w$ are the different values of X_k. For example if the individual-level covariate X_1 is town of residence, then the pair-level covariate Z_1 could indicate if the individuals live in the same town, if they live in neighboring towns, or if they live in towns that are not neighboring.

Naive Bayes

In order to estimate the probability that cases i and j are linked by direct transmission, $p(i \to j)$, we used a classification technique called naive Bayes. Naive Bayes has been used in many applications in machine learning and public health including text classification [35], software defect testing [36], and creation of genetic risk scores [37]. The method utilizes Bayes rule to estimate the probability of an outcome given a set of covariates from the observed frequencies in the training set.

For our method, the outcome variable is whether or not the pair is linked by direct transmission. We call the outcome variable L_{ij}, which takes the value 1 if case i infected case j and 0 otherwise. We know the probable value of this variable for case pairs in the training set based on pathogen WGS or contact investigation data and our method predicts the probability that $L_{ij} = 1$ for the rest of the case pairs.

We first use the training set to calculate, $P(Z_k = z_k|L = l)$, the probability that covariate Z_k is equal to z_k given link status, l, ($l = 1$ for linked pairs and $l = 0$ for non-linked pairs) for each covariate $k \in \{1, 2, \dots, p\}$ using:

$$P(Z_k = z_k|L = l) = \frac{\sum_{i,j} \mathbb{1}\{L_{ij} = l, Z_{kij} = z_k\} + 1}{\sum_{i,j} \mathbb{1}\{L_{ij} = l\} + n_k}. \tag{3}$$

Here the function, $\mathbb{1}$, is the indicator function which takes the value 1, if the input is true and 0 if it is false and n_k is the number of levels of Z_k for $k \in \{1, 2, \dots, p\}$. Therefore, the numerator, $\sum_{i,j} \mathbb{1}\{L_{ij} = l, Z_{kij} = z_k\}$, counts the number of instances that a pair, i, j has linked status, l, and covariate value z_k for covariate Z_k, $k \in \{1, 2, \dots, p\}$. Because sparse data can result in zero probabilities, we also apply a common technique of adding 1 to each cell called Laplace smoothing [35]. Then we use the training set to calculate $P(L = l)$, the prior probability of link status for $l \in \{1, 0\}$.

$$P(L = l) = \frac{\sum_{i,j} \mathbb{1}\{L_{ij} = l\} + 1}{N + 2}. \tag{4}$$

We now use Bayes rule to calculate the predicted probability that case i infected case j, $p(i \to j)$ for all pairs in the prediction set as:

$$\begin{aligned} p(i \to j) &= P(L_{ij} = 1|Z_{1ij} = z_1, \dots, Z_{pij} = z_p) \\ &= \frac{P(Z_1 = z_1, \dots, Z_p = z_p|L = 1)P(L = 1)}{P(Z_1 = z_1, \dots, Z_p = z_p)} \\ &= \frac{\prod_{k=1}^{p} P(Z_k = z_k|L = 1)P(L = 1)}{\prod_{k=1}^{p} P(Z_k = z_k|L = 1)P(L = 1) + \prod_{k=1}^{p} P(Z_k = z_k|L = 0)P(L = 0)}. \end{aligned} \tag{5}$$

We can calculate the conditional probability of all covariate values given link status, $P(Z_1 = z_1, \dots, Z_p = z_p|L = l)$, as the product of the conditional probabilities of each covariate, $P(Z_k = z_k|L = l)$ for $k \in \{1, 2, \dots, p\}$ in equation 5, under the assumption that covariates are conditionally independent. Note that the ordered nature of the pair dataset implies that if case j was observed before case i, then $p(i \to j) = 0$.

Scaling Probabilities

In order to compare probabilities obtained with different datasets, covariates, and ways of defining transmission events in the training set, the estimated probabilities are scaled so that they represent the relative likelihood that case j has been infected by case i as opposed to any other sampled case as:

$$p(i \to j)^S = \frac{p(i \to j)}{\sum_{m \neq j} p(m \to j)}. \tag{6}$$

This scaled probability, $p(i \to j)^S$, will henceforward be referred to as the relative transmission probability.

Training Dataset Construction

The implementation of naive Bayes involves using a training set in which the outcome is known to train the model and a prediction set in which the probabilities are estimated. In many applications these are two separate datasets. In the infectious disease applications we are considering though, we never know the true infector. Therefore in our training set the outcome, direct transmission between case pairs, is not known with certainty. However, many studies have more detailed information on a subset of cases, such as pathogen WGS and/or detailed contact investigation data which allow for the identification of probable linked and un-linked case pairs. We can use one or both of these sources of information to create a training set of probable transmission events. The rest of the cases which do not have these sources of information, are only included in the prediction set.

Because the outcomes in our training set represent probable and not certain transmission events, we want to estimate the transmission probability of these case pairs as well as the other pairs which do not have pathogen WGS or contact investigation data. Therefore, we use an iterative estimation procedure to ensure that each pair has a turn in the prediction set. We split our training set into n subsets called folds where n - 1 folds are used to train the model and the remaining fold is included in the prediction set. In our situation, this procedure is complicated by the fact that we assume that each case has only one true infector which presents two problems: 1) since the method of defining transmission events in the training set is not perfect, there could be multiple possible links for each infectee, and 2) once we denote a case pair as linked in the training set, all pairs that share that infectee have a zero-probability of being linked.

To solve the first problem, we create the links in the training set by randomly choosing one of the possible infectors to be designated the linked case pair and then repeated this selection multiple times to capture the uncertainty around the true infector. To solve the second problem, when a pair is designated as a link in the training set, all pairs that shared the infectee with the link are also included in the training set as non-links. Then when that linked pair is in the prediction dataset so are all of the other pairs involving the infectee, so the probability of all pairs involving that infectee can be estimated.

In order to capture the valuable information provided by the pathogen WGS or contact investigation data used to define probable transmission events in the training set, when a pair is included in the training set, the predicted probability for that run is set to 1 if the pair is linked and 0 if the pair is not linked. The following algorithm describes the creation of the training dataset and the iterative estimation procedure which can be visualized in Figure 1.

1. Create a dataset of possible training case pairs by subsetting the dataset to only the pairs involving individuals that have information on the variables used to define probable links in the training set.
2. Randomly select one infector for each infectee and designate those pairs as “linked” ($L_{ij} = 1$).
3. Temporarily remove all pairs that share an infectee with the linked pairs defined in step 2 from the dataset.
4. Designate all remaining un-linked pairs as not linked ($L_{ij} = 0$).
5. Split this dataset of possible training pairs into n folds (we used 10): n-1 for training and 1 for prediction.
    (a) Reserve 1 fold for prediction and combine with all of the pairs not in the training set.

   (b) For all linked pairs in the n-1 training folds, move all other pairs involving the infectee from the prediction set to the training set as non-links ($L_{ij} = 0$). This is the final training set for this iteration.

   (c) Set the predicted probabilities for training set pairs to 1 for links and 0 for non-links.

   (d) Use the training set to train the model and calculate predicted probabilities in the prediction set.

   (e) Repeat (a)-(d) n times so that each fold has a turn in the prediction set.

6. Repeat steps 2-5 multiple times (we used 10) to allow for different possible infectors to be designated the true infector.

7. Average over all the predicted probabilities for each pair.

8. Scale the resulting probabilities using equation 6 to obtain the relative transmission probabilities for all case pairs.

Reproductive Number Estimation

To estimate the reproductive number we use the approach described by Wallinga and Teunis [4]. These authors calculated the relative likelihood that each case had been infected by all other cases based on the serial interval distribution. They then calculated the effective reproductive number (R_i) for each case by summing up the scaled probabilities for all possible infectees as:

$$R_i = \sum_{m \neq i} p(i \rightarrow m)^S. \tag{7}$$

We use the same equation, but instead of deriving the probabilities based on the serial interval, we use the probabilities derived from our naive Bayes approach. This method assumes that all cases are sampled and that the outbreak is completed.

By averaging the individual reproductive numbers for all cases infected each month, we obtain estimates of the monthly effective reproductive number (R_t). Finally, we average the monthly reproductive numbers for the stable portion of the outbreak to estimate the average effective reproductive number over the study period ($\bar{R_t}$).

We can estimate confidence intervals for R_t and $\bar{R_t}$ using parametric bootstrapping. The R_i values are re-sampled 100 times using the estimated probabilities, $p(i \rightarrow k)^S$, for all k according to their distribution:

$$R_i \sim \sum_{m \neq i} Bernoulli(p(i \rightarrow m)^S) \tag{8}$$

as detailed in Wallinga and Teunis [4]. For each re-sampling we calculate the R_t values at each month and average them to estimate $\bar{R_t}$. We then use the distributions of the estimated values of R_t and $\bar{R_t}$ to derive 95% bootstrap confidence intervals:

$$LowerBound = \hat{R} - Q_{\tilde{R}}(1 - \alpha/2) - \hat{R} \tag{9}$$

$$UpperBound = \hat{R} - Q_{\tilde{R}}(\alpha/2) - \hat{R} \tag{10}$$

where R can be either $\bar{R_t}$ or R_t and $Q_{\tilde{R}}$ is the quantile function of the bootstrap estimates of R.

Simulation Study

We assess our method by applying it to simulated outbreaks. We simulate 1000 outbreaks using the **simulateOutbreak** function and generate the phylogenetic trees for those outbreaks using the **phlyoFromPTree**, function both from *TransPhylo* v1.0 [23]. Then we use the **simSeq** function in *phagnorn* v2.5.3 [38] to generate genetic sequences corresponding to the phylogenetic tree. This method was used in Stimson et al. [39] for a similar purpose, though we extend it to include multiple transmission chains.

Each outbreak is composed of numerous transmission chains of at least two cases which are simulated iteratively until the total sample size exceeded 500. The different transmission chains last for 20 years but have start and end points that vary in time. We simulate each transmission chain with the same outbreak parameters: a mean reproductive number ($\bar{R}_t$) of 1.2, a serial interval distribution of gamma(shape = 1.05, scale = 2.0), and an effective population size times generation time (N_{eg}) of 0.25. We also include a delay between infection and sampling using the same distribution as the serial interval. The serial interval we use was estimated from a TB household contact story in Brazil (Ma et al, under review). We then simulate representative pathogen genomes for each case. For each case we also simulated four different covariates representing clinical and demographic variables to inform the model. We also included the time between infection dates for each case pair into the model categorized as follows: less than 1 year, 1 to 2 years, 2 to 3 years, 3 to 4 years, 4 to 5 years, and more than 5 years.

Although pathogen genomes are thousands to millions of base-pairs long, we simulate a 300 base pair genome where each transmission chain starts with a unique set of base pairs to allow for genetic diversity across different strains. We aim to replicate a slow mutating pathogen such as TB which mutates at a rate of around 0.5 single-nucleotide polymorphisms (SNPs) per genome per year [11]. With this mutation rate, over the course of one 20 year transmission chain, very few mutations will accrue thus allowing a smaller genome to provide a good proxy for the full genome. The shortened genome length we simulate is meant to represent the locations that differ amongst cases sampled as part of one outbreak instead of the full genome. We also performed a sensitivity analysis which showed that the SNP distance for a fixed mutation rate did not notably change as genome length increased (Figure 2). It is possible that this approach underestimates the true SNP distance distribution which makes it a conservative representation of the true SNP distance distribution using the full pathogen genome.

We compare the performance of our method when we train the model using probable transmission events defined by SNP distances between case pairs verses training the model with truly linked and un-linked pairs case pairs (Table 1). When we train the model using SNP distance, case pairs with less than 2 SNPs are considered links and those with more than 12 SNPs are considered non-links in the training set. Pairs with between 2 and 12 SNPs are considered indeterminate and thus are not included in the training set. To simulate real scenarios when only a fraction of the cases have the discriminating information necessary to define probable transmission events, we randomly select a subset of 50% of all cases to make up the training set.

Table 1: Description of simulation scenarios

Simulation Scenarios
Model trained on true links
Model trained on SNP distance links
Correct serial interval - gamma(1.05, scale = 2.0)
Wide serial interval - gamma(1.3, scale = 3.3)
Narrow serial interval - gamma(0.54, scale = 1.9)
Random probabilities

For comparison we also compute the relative transmission probabilities based on the serial interval used to simulate the data and the time between infection dates of each pair. We also compute probabilities using two mis-specified serial intervals, one that is too wide and one that is too narrow (Table 1). This comparison is motivated by the Wallinga and Teunis [4] method for calculating the reproductive number in which the transmission probabilities are estimated from the serial interval distribution. The wide and narrow serial intervals represent the prior and posterior distributions used by Didelot et al. [23] in the analysis of the same TB outbreak in Hamburg that we analyze below. All of these serial intervals are shifted by 90 days because a serial interval of less than 3 months is not possible for TB, our motivating example. We also assign the probabilities randomly from a Uniform(0, 1) distribution as a negative control.

We estimate the monthly effective reproductive number, R_t for all of the scenarios. To estimate the average effective reproductive number, $\bar{R}_t$, we average the monthly values for all months excluding the first 10% and last 30% of months for each outbreak due to the unobserved infectors or infectees at the beginning and the end of the sampling period.

In order to determine what proportion of cases need to be included in the training set to achieve good performance, we also perform a sensitivity analysis where we simulate 300 outbreaks with sample sizes ranging from 50-1000 cases. For each outbreak we perform our method assuming that a random subset of 10% to 100% of the cases can be included in the training set. We assess how changing the proportion of cases sampled affects the performance of the model performance and the estimation of the reproductive number depending on the sample size.

Application to a TB Outbreak in Hamburg

Finally, we apply our method to a small TB outbreak in Hamburg and Schleswig-Holstein, Germany that Roetzer et al. [10] also analyzed. The outbreak included 86 individuals from the largest strain cluster in a long-term surveillance study conducted by the health departments in these cities. The dataset includes pathogen WGS data for all individuals as well as clinical, demographic, and social risk factor data. Furthermore a subset of these individuals were involved in contact investigations performed by the local health authorities.

We used two different methods of defining probable links in the training set: 1) SNP distances and 2) contact investigation. When the model was trained with SNP distances, pairs with < 2 SNPs between them were considered linked, pairs with > 12 SNPs between them were considered non-links, and pairs with between 2 and 12 SNPs were considered indeterminate and were not included in training set. When the model was trained using contact investigation, pairs that had confirmed contact with each other were considered linked, pairs without confirmed contact with each other were considered non-linked, and cases who did not undergo contact investigation were not included in the training set. We also calculated the relative transmission probabilities randomly and using the same serial intervals as the simulation study for comparison.

Results

Simulation Study

We created 1000 simulated outbreaks composed of multiple transmission chains with a total of at least 500 cases. The sample sizes ranged from 500 to 1178 (median: 545) and each outbreak consisted of between 2 and 39 (median: 14) individual transmission chains with 2 to 846 (median: 9) cases each.

Figure 3 shows the distribution of relative transmission probabilities in one of the 1000 outbreaks for the different scenarios comparing truly linked and non-linked case pairs. For both ways of defining links in the training set, our method estimated relative transmission probabilities of less than 0.005 for most of the non-linked pairs (92% training using true links and 87% training using SNP distance). We also found that with either way of defining the training set, our method assigns more than 80% of truly linked case pairs higher probabilities than the serial interval method (Figure 4).

In order to assess how well the relative transmission probabilities could classify case pairs as linked and non-linked across all 1000 simulated outbreaks, we calculated the area under the receiver operating curve (AUC) for each simulation. To determine how well our method performed in identifying the true infector, we evaluated how the relative transmission probability of the true infector ranked compared to all possible infectors.

Table 2: Performance metrics for relative transmission probabilities over 1000 simulations

Scenario	AUC	Correct	Top 5%	Top 10%	Top 25%	Top 50%
	Percent (SD)	Percent (SD)	Percent (SD)	Percent (SD)	Percent (SD)	Percent (SD)
Gold Std: Truth	96.2 (0.7)	34.1 (3.1)	73.2 (4.7)	83.7 (3.7)	94.4 (2.0)	98.9 (0.7)
Gold Std: SNP Distance	95.1 (1.0)	29.1 (2.7)	73.9 (4.3)	83.6 (3.6)	94.0 (2.0)	98.7 (0.8)
Correct Serial Interval	87.5 (1.9)	2.5 (1.0)	34.7 (6.4)	53.7 (7.4)	81.6 (5.5)	96.3 (1.9)
Wide Serial Interval	84.8 (2.2)	1.7 (0.8)	26.4 (6.1)	44.7 (8.2)	76.4 (5.5)	95.3 (2.5)
Narrow Serial Interval	87.2 (1.9)	2.5 (1.0)	34.7 (6.4)	53.7 (7.4)	81.6 (5.5)	96.3 (1.9)
Random Probabilities	60.6 (1.1)	0.4 (0.3)	5.2 (1.1)	10.2 (1.5)	25.2 (2.2)	50.1 (2.5)

The average AUC when the model was trained with true links was 96% (standard deviation [SD] 0.7) compared to 95% (SD 1.0) when the model was trained using SNP distances. The probability of the true infector was ranked in the top 25% of all possible source cases on average 94% (SD 2.0) of the time both when the model was trained with true links and links determined by SNP distances. These results show that our estimated probabilities can distinguish between linked and non-linked pairs even when we train the model with only probable instead of certain transmission events. The probabilities calculated by our method outperformed probabilities obtained using any of the serial intervals in classification ability and especially in finding the true infector (Figure 5, Table 2).

Figure 6 and Table 3 show the estimates $\bar{R}_t$ for each of the different scenarios compared to the value of 1.2 used to simulate the outbreaks. Our method estimated the reproductive number accurately as did the method using the correct serial interval. However, when incorrect serial intervals were used, the estimates of $\bar{R}_t$ were either too high or too low depending on whether the serial interval was too wide or too narrow respectively.

Table 3: Average effective reproductive number for different simulation scenarios

Scenario	$\bar{R}_t$, mean (SD)
True Transmission	1.18 (0.09)
SNP Distance	1.21 (0.11)
Correct Serial Interval	1.18 (0.09)
Wide Serial Interval	1.30 (0.15)
Narrow Serial Interval	1.11 (0.05)
Random Probabilities	1.50 (0.26)

In our sensitivity analysis of the proportion of cases needed to train the model, as expected, we found that the performance improves and the variability in the metrics decreases as the proportion of cases in the training set increases (Figure 7). However, if the sample size is at least 500 then only 10% of all cases are needed to train the model to obtain good performance. If the sample size is between 200 and 500, training the model with 20% of cases results in good performance. For smaller outbreaks, at least 40% of the cases need to be in the training set to obtain the same performance level (Figure 8). When there is a small proportion of cases in the dataset, our method over-estimates the average effective reproductive number slightly. The estimates get increasingly accurate as the training dataset increases and as long as around 40% of cases are included in the training dataset the estimates are close to the true value (Figure 9).

Application to a TB Outbreak in Hamburg

We applied our method to a small TB outbreak in Hamburg and Schleswig-Holstein, Germany that Roetzer et al. [10] also analyzed. The outbreak included 86 individuals from the largest strain cluster in a long-term surveillance study conducted by the health departments in these cities. The cases were sampled from 1997 to 2010 with two spikes in reporting (Figure 10). There were 62 (72%) cases from Hamburg and 24 (23%) cases from Schleswig-Holstein. The median age was 45 (range: 2-85) and 70 (81%) of the cases were men. Details on the other clinical and demographic characteristics can be seen in Table 4.

These 86 cases resulted in 3633 possible ordered case pairs where the possible infector was observed before the possible infectee. These pairs were separated by between 0 and 20 SNPs, with a median of 4 SNPs between them. Of the 86 individuals, 31 (36%) were part of contact investigations performed by the local health authorities resulting in 51 case pairs with confirmed contact with each other. All individual-level covariates were transformed into pair-level covariates as detailed in Table 6. We used these covariates to estimate the relative transmission probability for all case pairs from this outbreak. We trained the model in two ways: using SNP distances or contact investigations.

Figure 11 shows heatmaps of all potential infectors for each infectee using our method compared to random probabilities (Figure 11A) and the correct serial interval (Figure 11B). When our method is applied, either with SNP distance defining links in the training set (Figure 11C) or with confirmed contact defining links (Figure 11D), the plots display more variation in the relative transmission probability across possible infectors than the serial interval or random scenarios. For some infectees there are infectors with a higher probability than all others in the row indicating the likely true infector. However, even for rows without a

Table 4: Individual-level demographic and clinical characteristics for the Hamburg outbreak

Covariate	Level	All Individuals
		(n = 86)
Nationality	Germany	66 (76.7%)
	Other	20 (23.3%)
Substance Abuse	No	33 (38.4%)
	Yes	53 (61.6%)
HIV	Negative	81 (94.2%)
	Positive	5 (5.8%)
Residence	Permanent residence	71 (82.6%)
	Homeless	15 (17.4%)
Association with milieu street drinking scene	No	21 (24.4%)
	Yes	65 (75.6%)
City	Hamburg	62 (72.1%)
	Schleswig-Holstein	24 (27.9%)
Age Group	$<$ 30 years old	10 (11.6%)
	30-39 years old	16 (18.6%)
	40-49 years old	28 (32.6%)
	50-59 years old	15 (17.4%)
	$\geq$ 60 years old	17 (19.8%)

clear single infector, many possible infectors have very low probabilities and can be eliminated as possible infectors.

Table 5: Average effective reproductive number for Hamburg outbreak by method

Scenario	$\bar{R}_t$, mean (SD)
Confirmed Contact	0.96 (0.73, 1.18)
SNP Distance	0.89 (0.67, 1.09)
Narrow Serial Interval	1.07 (0.82, 1.31)
Medium Serial Interval	0.98 (0.68, 1.23)
Wide Serial Interval	0.88 (0.64, 1.11)
Random Probabilities	0.78 (0.54, 1.00)

For all of the methods, we calculated R_t per month and $\bar{R}_t$ with bootstrap confidence intervals. Since this was a smaller time frame than our simulations, when estimating $\bar{R}_t$ we only removed the last 10% of the months. The monthly R_t values for each scenario can be seen in Figure 12. All of the methods except random probabilities show spikes in monthly reproductive numbers right around the second peak in case counts (Figure 12), but to different degrees.

The estimates of $\bar{R}_t$ for all of the methods were approximately 1 with some variation (Table 5, Figure 13). When the probabilities were estimated using contact investigations to train the model and using the medium serial interval, we estimated $\bar{R}_t$ to be 1.0 (95% confidence interval [CI] 0.73-1.18 and 0.67-1.23, respectively). When the probabilities were estimated using the SNP distances to train the model and using the wide serial interval we estimated $\bar{R}_t$ to be 0.9 (95% CI 0.67-1.09 and 0.64-1.11, respectively). Using random probabilities resulted in a lower estimate of 0.8 (95% CI 0.54-1.00) and using the narrow serial interval resulted in a higher estimate of 1.1 (95% CI 0.81-1.31).

Discussion

We have developed a method to estimate the relative transmission probability between pairs of infectious disease cases using clinical, demographic, geographic, and genetic characteristics that accurately distinguishes between linked and non-linked case pairs. These probabilities can be used to explore possible transmission chains, rule out transmission events, and estimate the reproductive number. Our method was assessed in a series of simulation studies as well as applied to a TB outbreak in Germany.

In simulation studies, we found that the probabilities were able to distinguish between linked pairs and non-linked case pairs with high accuracy. Using a SNP distance proxy for transmission to train the model, we achieved a classification accuracy of 95% and 94% of the time the true infector was assigned a probability in the top 25% of all possible infectors. Our method outperformed the serial interval method of estimating probabilities in all classification and identification metrics.

We also showed that our method accurately estimates the average effective reproductive number in simulated outbreaks. Although probabilities derived from true serial interval also correctly estimated the reproductive number, probabilities derived from incorrect serial intervals did not. Therefore, in absence of a reliable serial interval estimate, the probabilities from our model better estimate the effective reproductive number than using the serial interval distribution. This result is important because the serial interval is difficult to estimate and may vary across different settings [7, 8, 40].

The Hamburg outbreak provided a way to evaluate our method using two different ways of defining probable transmission events in the training set: SNP distance and contact investigation. We found that either method allowed for the elimination of many transmission links (Figure 11). The two ways of training the model did not produce identical results which is expected because the two models are trained on different ways of defining probable transmission events, neither of which perfectly capture the truth. Using contact investigation to train the model is more discriminating than using SNP distance because the cases are known to have interacted, but it also could also miss links from unknown contacts. Using SNP distances to train the model will result in fewer missed links, but it could connect cases who never interacted. The two methods result in similar estimates for the $\bar{R}_t$ and we hypothesize that the true reproductive number for *M. tuberculosis* in this context lies in between these two estimates (0.9 - 1.0).

Most established methods for exploring transmission within an outbreak focus on either identifying clusters of recent transmission [11, 14, 39, 41–44], recreating the transmission chain [12, 13, 15–22, 45], or attempting to identify the true source case [46–50]. When it comes to estimating transmission parameters, simply knowing clusters does not provide enough information and identifying the true infector is often impossible in practical applications. The strength of our method is that it does not aim to find the one true infector but estimates the relative probability of transmission for all sampled infectors. This gives our method broad applicability. It can be used to find potential true infectors - pairs with very high probabilities or identify clusters of transmission - groups of pairs with high probabilities. The probabilistic nature of the estimates can then be used to further estimate transmission parameters incorporating the uncertainty around the true infector.

Other methods to estimate the transmission probability between cases either only use genetic data [23] or require prior knowledge of the relationship between the covariates and transmission [51]. Our method, on the other hand, harnesses many different sources of information about the characteristics of individual cases and the differences between cases in a way that does not presuppose any relationship between these factors and transmission.

Additionally our method is built upon naive Bayes, a simple but powerful machine learning tool that has been used in a large variety of applications [35–37, 52, 53]. It is straightforward to implement and computationally efficient. Although traditionally a naive Bayes model is trained with a training set of true events, this will never be available for real-life infectious disease applications; however, WGS data for a subset of cases may be readily available in some outbreak settings. We have shown that our method performs almost as well when SNP distance is used as a proxy for transmission to train the model as when true transmission is used as to define links in the training set.

The structure of a training and prediction set also means that not all cases need to have the highly discriminatory information such as contact investigation or pathogen WGS data to estimate the relative transmission probabilities. This is a useful feature because existing datasets often have rich demographic, clinical, and spatial data but lack detailed contact investigation or pathogen WGS data because both of these data sources require significant time and resources to obtain. As long as a subset of cases, between 10% and 40% depending on the sample size, has this information, our method can be used to infer transmission patterns among the remaining cases as well.

Our method, as with any statistical model, makes certain assumptions which may or may not be universally appropriate. First of all, naive Bayes assumes that the covariates are independent when conditioning on the outcome, may not be realistic. However, numerous papers have shown that naive Bayes still performs extremely well even when this assumption is violated [53–56]. Furthermore, many extensions of naive Bayes have been developed which relax this assumption [35, 57, 58], which could be easily integrated into our method.

The Wallinga and Teunis [4] approach for estimating the effective reproductive number, which we applied, assumes that every case has been infected by someone that has been sampled . These authors and others found that simulations incorporating random incomplete reporting did not substantially decrease the accuracy of their reproductive number estimates, so we do not expect this to be an issue with our analysis [4, 59]. Our probability estimates themselves do not assume all cases in an outbreak are sampled because we are simply estimating how much more likely it is that one case was infected by another than by any other sampled case. However, our method could be affected by biased sampling - either because only certain types of cases are observed or only certain types of cases have the information needed to define probable links in the training set. Future work could more fully examine the effect of biased reporting and biased training sets.

Finally, our method has the same limitations as many other infectious disease analytical approaches in that it assumes that cases were infected in the same order that they were observed [44, 60]. This is not a strong assumption for a disease that has clear symptoms and a short latent period where most patients seek care shortly after developing symptoms. However, in diseases such as TB, where there can be months to years between infection and development of disease (if this ever occurs), and where delay in seeking care and getting properly diagnosed can be substantial [61, 62] it is possible that cases present to care in a different order than they were infected. Although this assumption is a known problem in infectious disease research, it is frequently made [46, 47] because relaxing it complicates models substantially.

We have developed a method to estimate the relative transmission probabilities between pairs of cases. Our method is flexible, using whatever sources of information are available without making any prior assumptions about the relationship between these covariates and transmission. The power of our method is that only a subset of cases (10-40% depending on the sample size) need to have pathogen WGS data or contact investigation data, allowing this method to be applied to a wide range of outbreak and surveillance datasets. These probabilities can be used to better understand the transmission dynamics of an outbreak by identifying possible transmission events and estimating transmission parameters. In a disease where determining transmission events can be extremely difficult, using probabilities of transmission between all possible cases provides a unique and powerful analysis tool.

Acknowledgments

This work was supported by the US National Institutes of Health (NIHGMS R01GM122876). SEV was also funded by the Interdisciplinary Training grant from the US National Institutes of Health (NIHGMS T32GM074905). HEJ was also funded by US National Institutes of Health (NIH K01AI102944). CRH and LFW were also funded by the Providence/Boston Center for AIDS Research (P30AI042853). CRH was supported by the Boston University/Rutgers Tuberculosis Research Unit (U19AI111276) and the U.S.-India Vaccine Action Program (VAP) Initiative on Tuberculosis (CRDF Global/NIAID). RSL holds a Fellowship from the Canadian Institutes of Health Research (MFE-152448). The content of the article is solely the responsibility of the authors and does not necessarily represent the views of the National Institute of Allergy and Infectious Disease or the Office of the Director, NIH. The funders had no role in the decision to publish this manuscript.

Table 6: Pair-level demographic and clinical characteristics for the Hamburg outbreak

Covariate	Level	All Pairs (n = 3633)
City	Same city	2148 (59.1%)
	Different city	1485 (40.9%)
Nationality	Both Germany	2129 (58.6%)
	Same foreign country	19 (0.5%)
	One Germany, one foreign country	1315 (36.2%)
	Different foreign countries	170 (4.7%)
Sex	Male to male	2401 (66.1%)
	Female to female	120 (3.3%)
	Male to female	757 (20.8%)
	Female to male	355 (9.8%)
Age group	Same age group	774 (21.3%)
	Different age group	2859 (78.7%)
Smear status	Infector smear-	1294 (35.6%)
	Infector smear+	2339 (64.4%)
HIV	Infector HIV-	3463 (47.8%)
	Infector HIV+	170 (4.7%)
Substance abuse	Both Yes	1372 (37.8%)
	Both No	526 (14.5%)
	Different	1735 (47.8%)
Residence	Both permanent	2062 (56.8%)
	Both homeless	105 (2.9%)
	Different	1055 (29.0%)
Association with milieu street drinking scene	Both Yes	2062 (56.8%)
	Both No	210 (5.8%)
	Different	1361 (37.5%)
Time difference	< 1 year	546 (15.0%)
	1-2 years	485 (13.3%)
	2-3 years	374 (10.3%)
	3-4 years	305 (8.4%)
	> 4 years	1923 (52.9%)
SNP distance	< 2 SNPs	796 (21.9%)
	2-12 SNPs	2452 (67.5%)
	> 12 SNPs	385 (10.6%)
Confirmed contact	Yes	51 (1.4%)
	No	408 (11.2%)
	Unknown	3174 (87.4%)

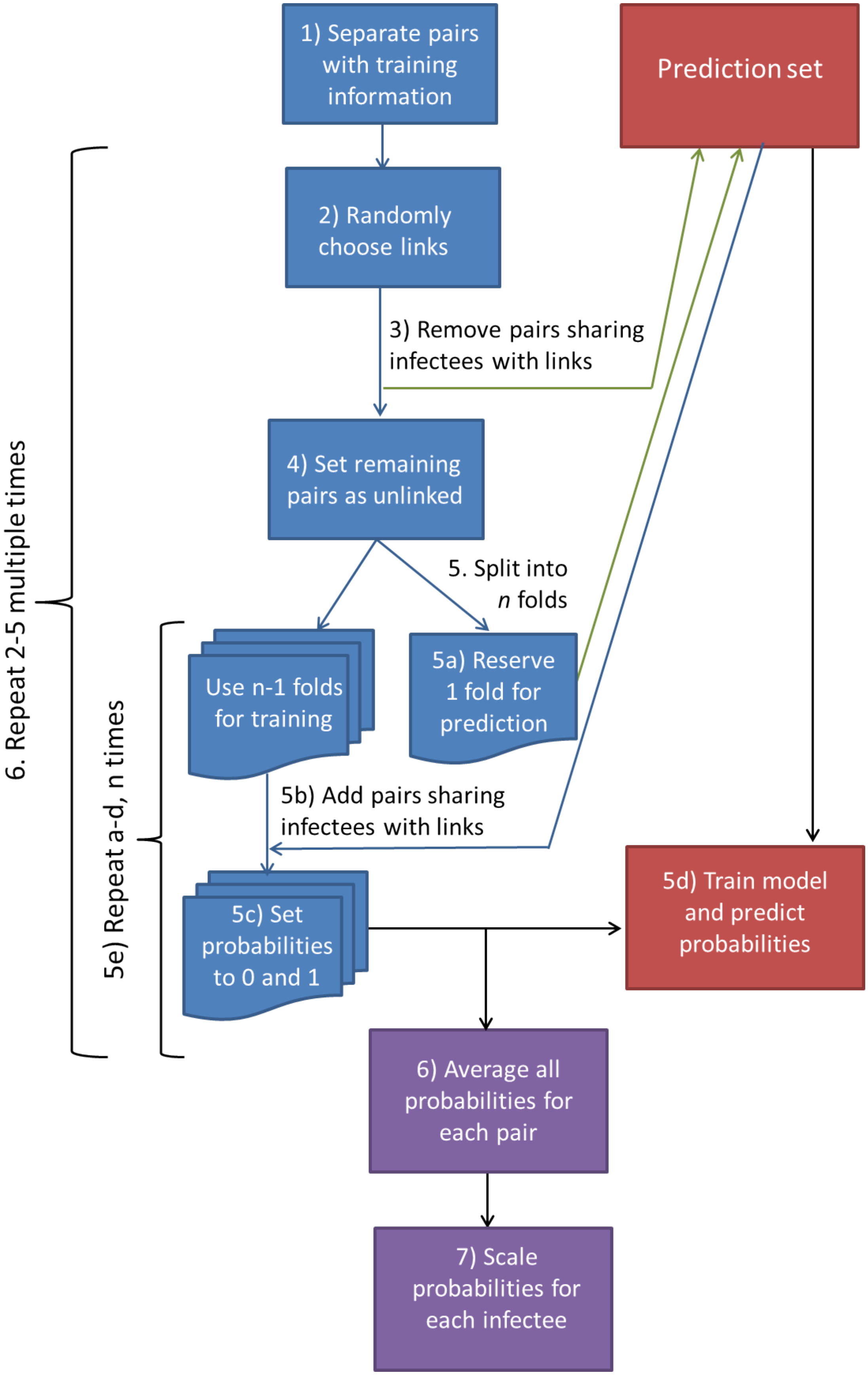
1) Separate pairs with training information
Prediction set
2) Randomly choose links
3) Remove pairs sharing infectees with links
4) Set remaining pairs as unlinked
5. Split into *n* folds
Use n-1 folds for training
5a) Reserve 1 fold for prediction
5b) Add pairs sharing infectees with links
5c) Set probabilities to 0 and 1
5d) Train model and predict probabilities
6) Average all probabilities for each pair
7) Scale probabilities for each infectee
6. Repeat 2-5 multiple times
5e) Repeat a-d, n times

Figure 1: Flow-chart depicting the algorithm we used to create the training dataset and the iterative procedure to estimate the relative transmission probabilities.

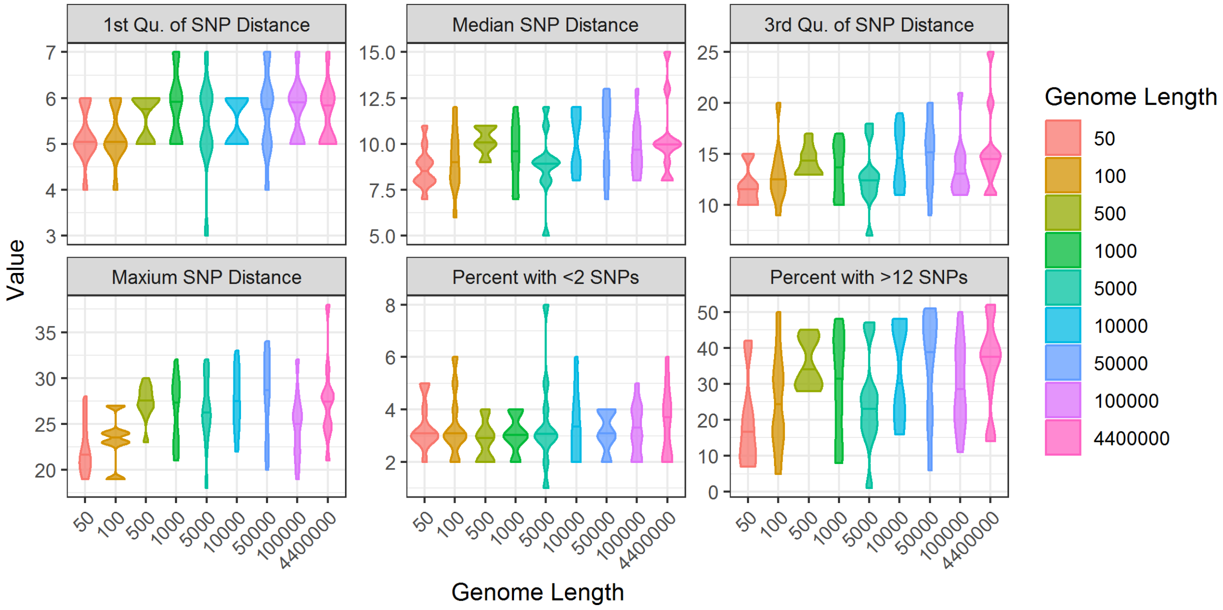

Figure 2: Violin plots representing the effect of changing genome length on the SNP distance distribution between case pairs. Pathogen genomes of various lengths from 50 to 4.4 million base pairs were simulated 100 times each for an outbreak of 200 cases. The resulting SNP distance matix for all case pairs was computed. The figure shows the relationship between genome length on the quartiles of the SNP distance distribution as well as the maximum SNP distance and the percent of pairs with less than 2 SNPs and more than 12 SNPs.

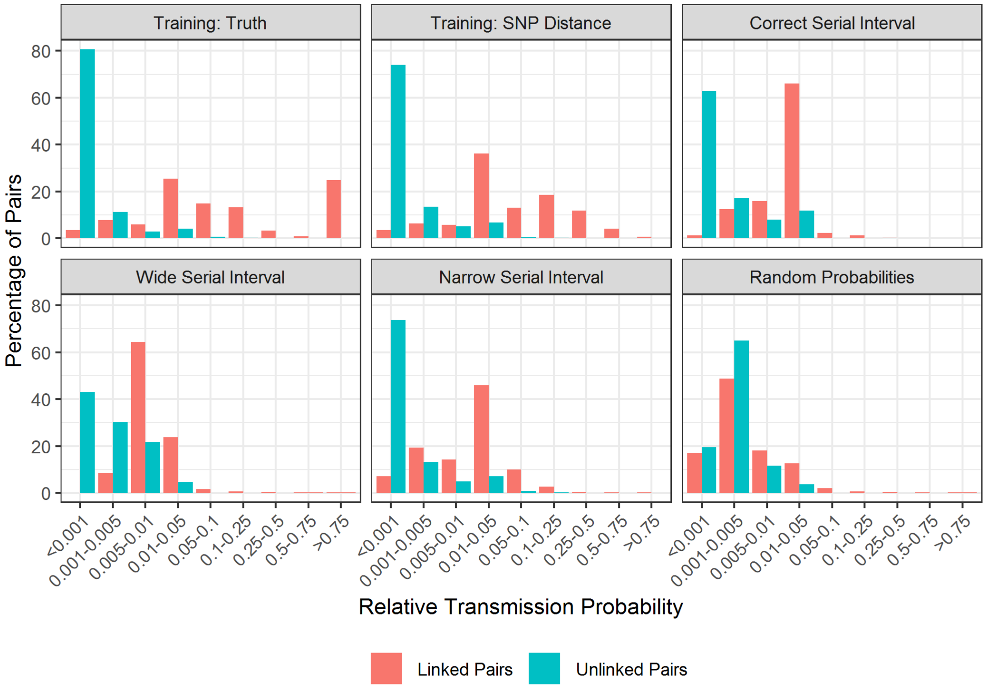

Figure 3: Distribution of the relative transmission probability – how much more likely is it that case i was infected by case j as opposed to any other sampled case – for linked case pairs and non-linked case pairs in one of the 1000 simulated outbreaks. Each panel shows a different method of calculating probabilities: our method with a training set of true links, our method with a training set of links defined by SNP distance, probabilities derived from the serial interval distribution used to simulate the outbreak: gamma(1.05, 2.0), probabilities derived from a serial interval distribution that is too wide: gamma(1.3, 3.3) and too narrow: gamma(0.54, 1.9), and random probabilities.

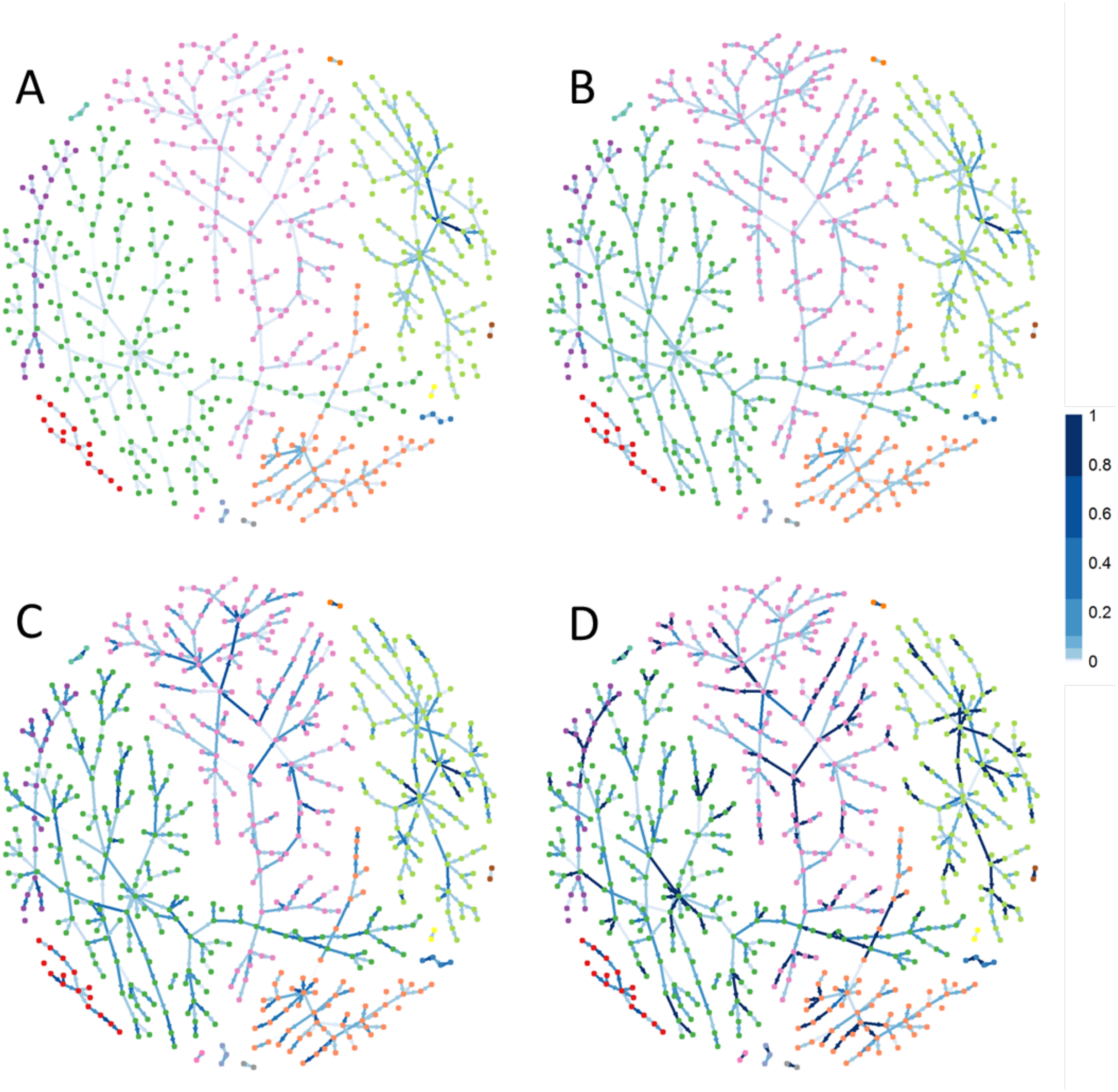

Figure 4: Network plots of the true transmission network in one of the 1000 simulated outbreaks. The nodes represent individual cases and are colored by transmission chain. The edges represent true transmission events and are colored based on the estimated relative transmission probability; the darker the color the higher the probabilities. A) Edges colored based on randomly assigned probabilities. B) Edges colored based on probabilities calculated by the correct serial interval: gamma(1.05, 2.0). C) Edges colored based on the probabilities calculated using our with a training set of links defined by SNP distance. D) Edges colored based on the probabilities calculated using our method with a training set of true links.

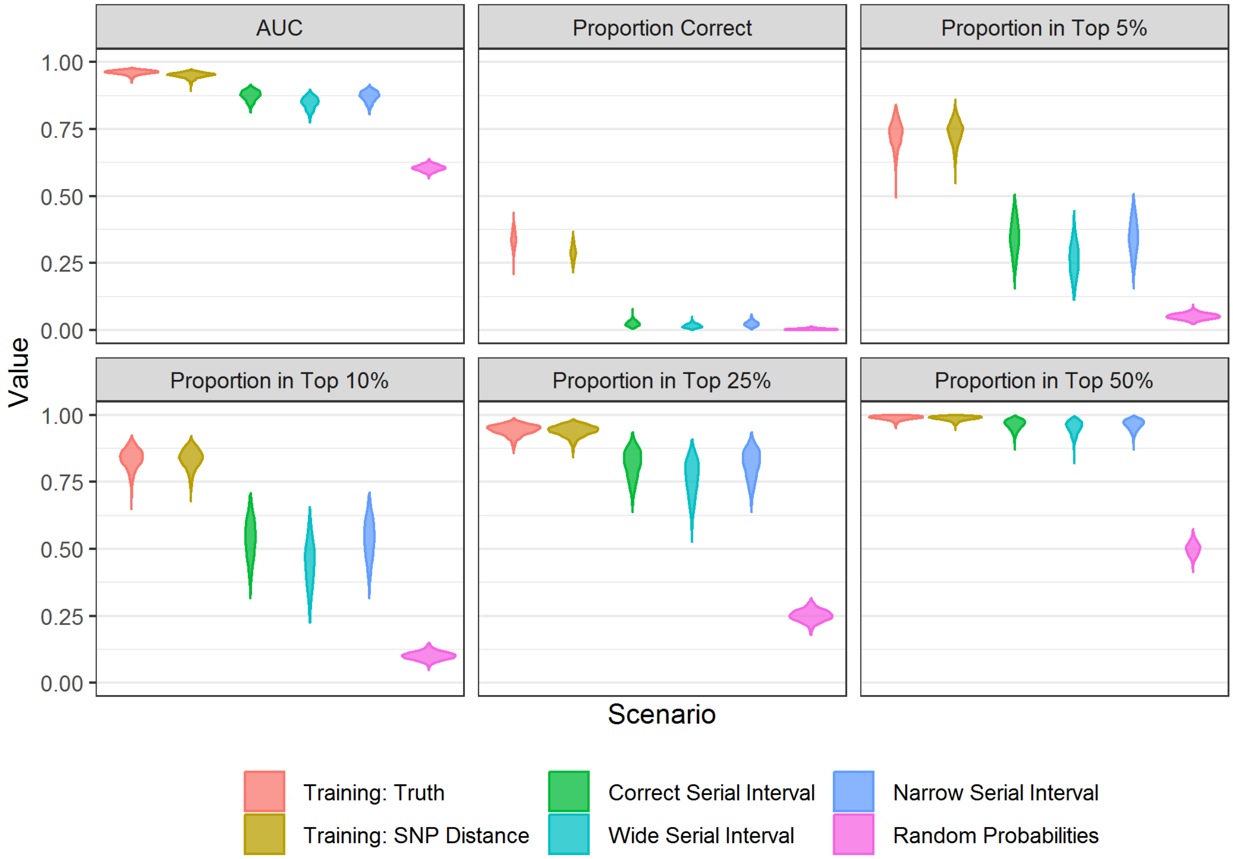

Figure 5: Violin plots of the performance metrics for the different scenarios across 1000 simulated outbreaks. The scenarios were: our method with a training set of true links, our method with a training set of links defined by SNP distance, probabilities derived from the serial interval distribution used to simulate the outbreak: gamma(1.05, 2.0), probabilities derived from a serial interval distribution that is too wide: gamma(1.3, 3.3) and too narrow: gamma(0.54, 1.9), and random probabilities. The metrics shown are the area under the receiver operating curve (AUC), the proportion of time the true infector was assigned the highest relative transmission probability (Proportion Correct), and the proportion of time the probability of the true infector was ranked in the top 5%, 10%, 25%, and 50% of all possible infectors.

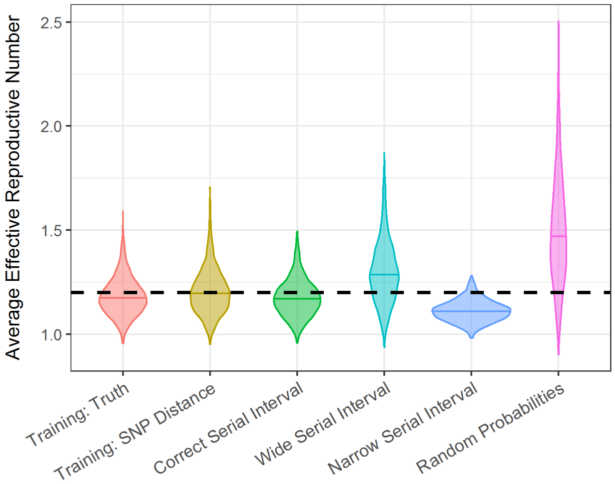

Figure 6: Violin plots of the distribution of the average effective reproductive number for different scenarios across 1000 simulated outbreaks. The dashed horizontal line indicates the true value of 1.2 that was used to simulate the outbreaks.

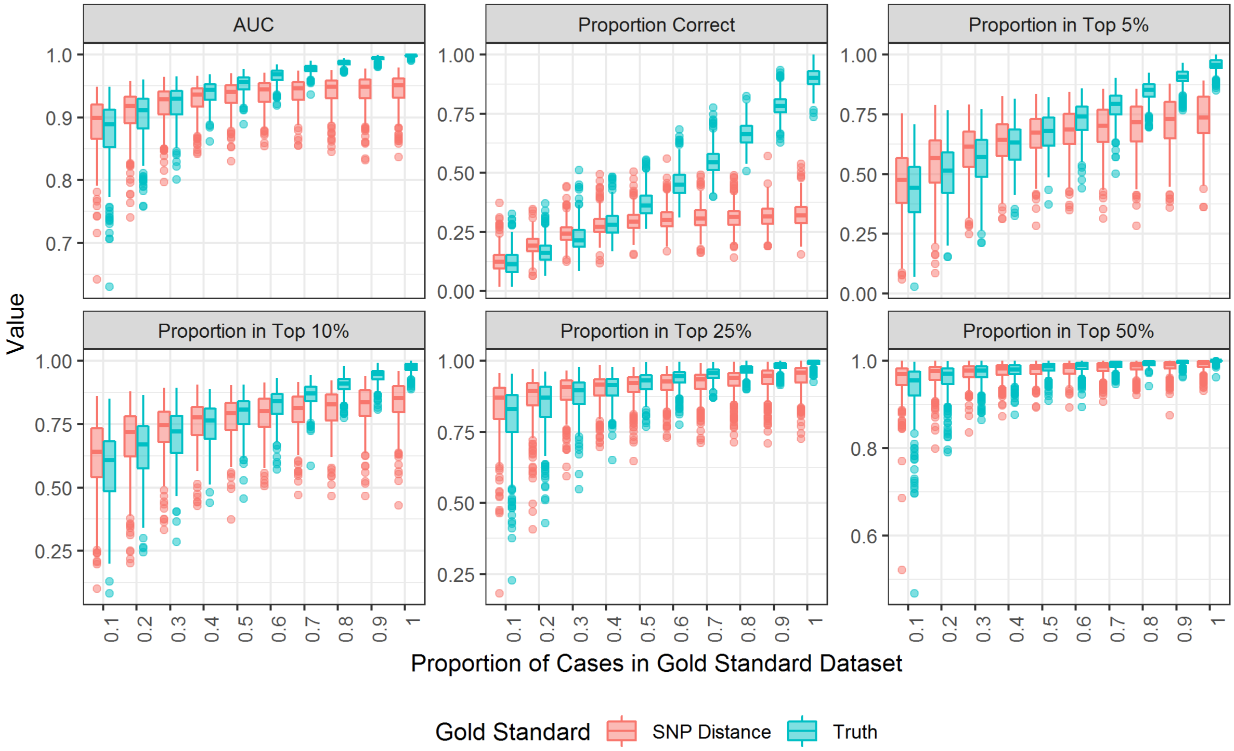

Figure 7: Boxplots of the performance metrics when varying the training set proportion in 300 simulated outbreaks with sample sizes varying from 50-1000. The plots are colored by the type of gold standard: SNP distance (red) or true transmission (blue). The metrics shown are the area under the receiver operating curve (AUC), the proportion of time the true infector was assigned the highest relative transmission probability (Proportion Correct), and the proportion of time the probability of the true infector was ranked in the top 5%, 10%, 25%, and 50% of all possible infectors.

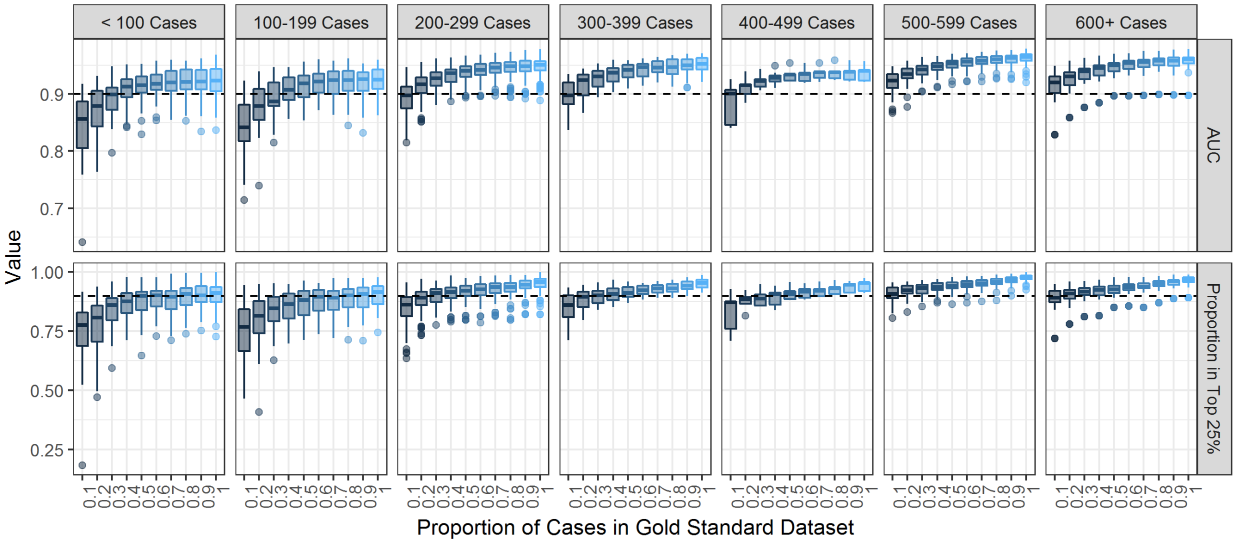

Figure 8: Boxplots of the performance metrics by training set proportion in 300 simulated outbreaks stratified by the total sample size of the outbreak. The metrics shown are the area under the receiver operating curve (AUC) and the proportion of time the relative transmission probability of the true source case was ranked in the top 25%. The dotted black line indicates a value of 90% on either metric.

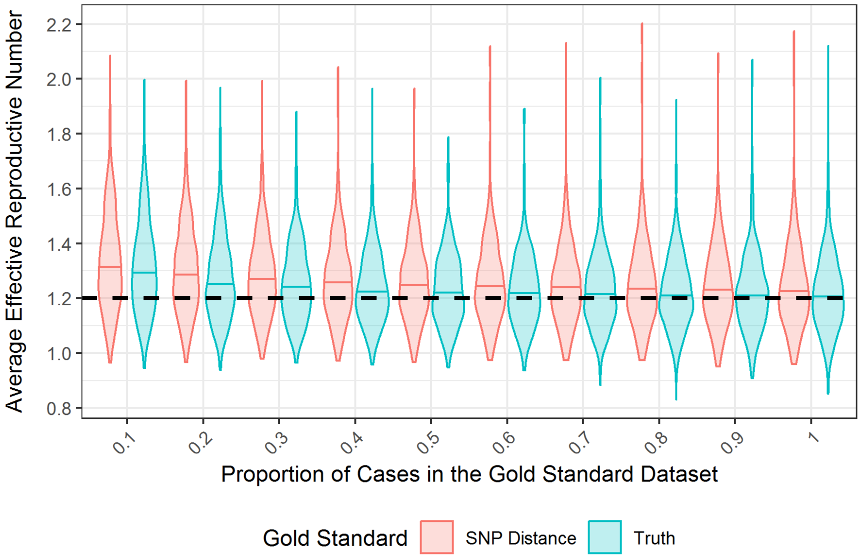

Figure 9: Violin plots of the distribution of the average effective reproductive number when varying the dataset proportion in 300 simulated outbreaks with sample sizes varying from 50-1000. The plots are colored by the way of defining links in the training set: SNP distance (red) or true transmission (blue). The dashed horizontal line indicates the true value of 1.2 that was used to simulate the outbreaks.

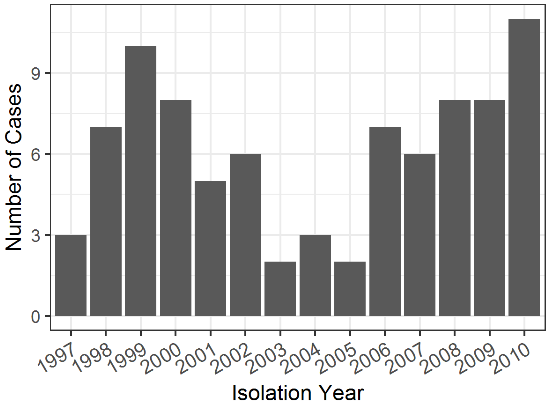

Figure 10: Case counts by year for the Hamburg outbreak described in Roetzer et al. [10]

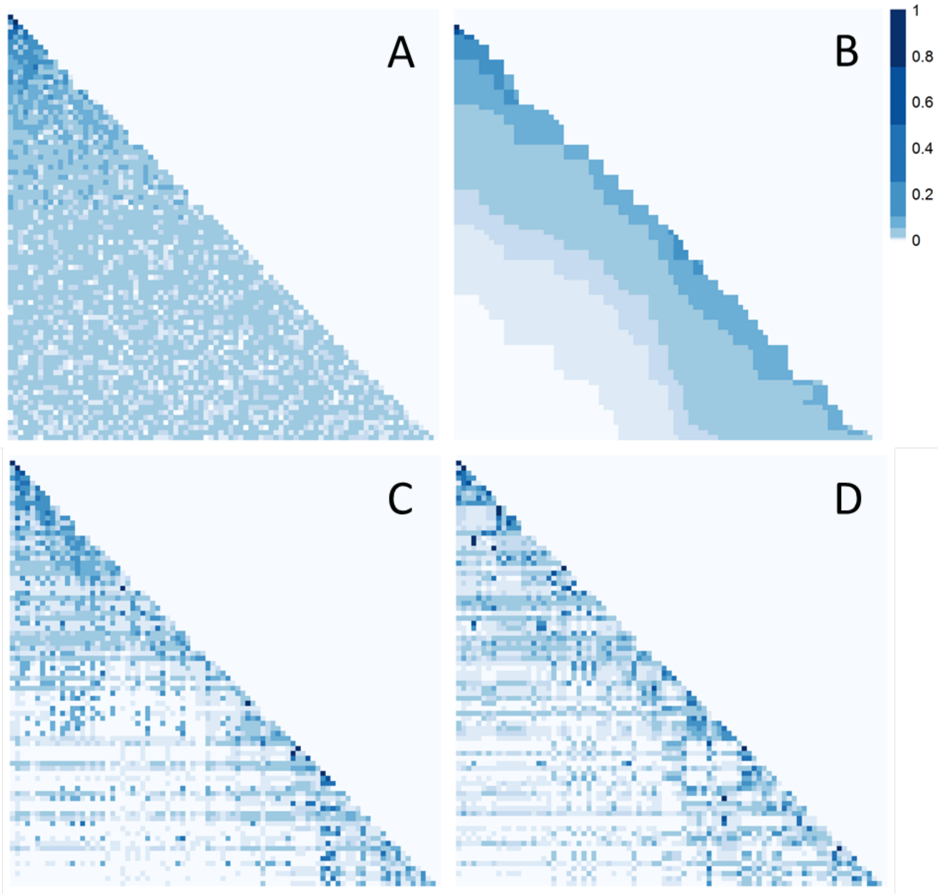

Figure 11: Heatmaps of the relative probabilities that each infectee (rows) was infected by each possible infector (columns) in the Hamburg TB outbreak. Darker squares represent higher the probabilities. The cases are ordered by infection date with the earliest cases on the top and to the left. Each panel shows the results from a different method of calculating probabilities: A) randomly assigned probabilities, B) probabilities calculated using a gamma(1.05, 2.0) serial interval distribution, C) probabilities calculated using our method and a training set with links based on SNP distance, and d) probabilities calculated using our method and a training set with links based on contact investigations.

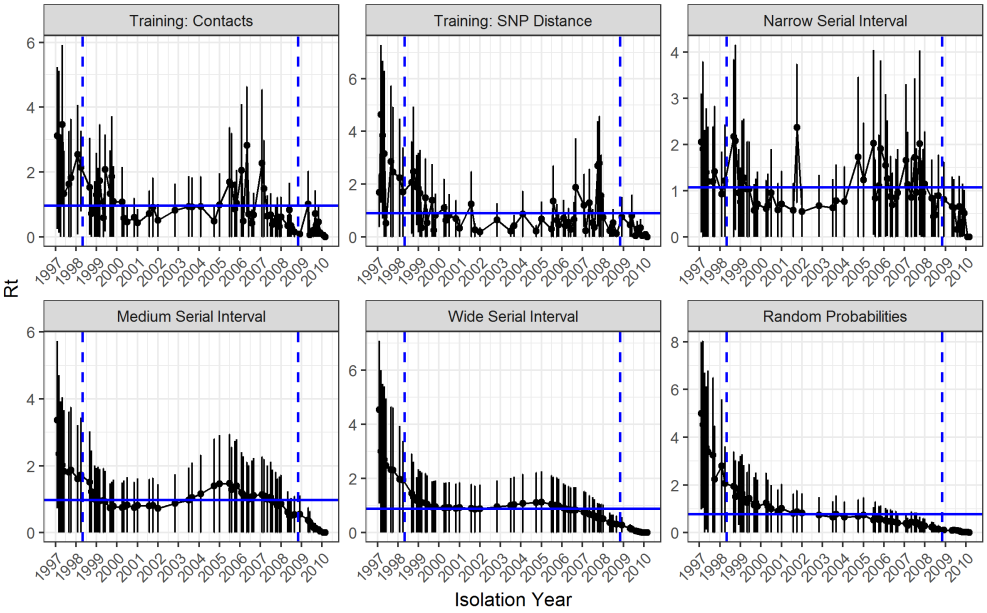

Figure 12: Monthly reproductive number over the course of the 14 years of the Hamburg TB outbreak estimated from the relative transmission probabilities with bootstrap confidence intervals. Each panel shows the results from a different method of calculating probabilities: our method and a training set with links based on contact investigation data; our method and a training set with links based on SNP distance; probabilities derived from narrow: gamma(0.54, 1.9), medium: gamma(1.05, 2.0), and wide: gamma(1.33, 3.0) serial interval distributions; and random probabilities. The months in between the dotted horizontal lines were averaged to find the average effective reproductive number for the scenario which is shown by the solid horizontal line.

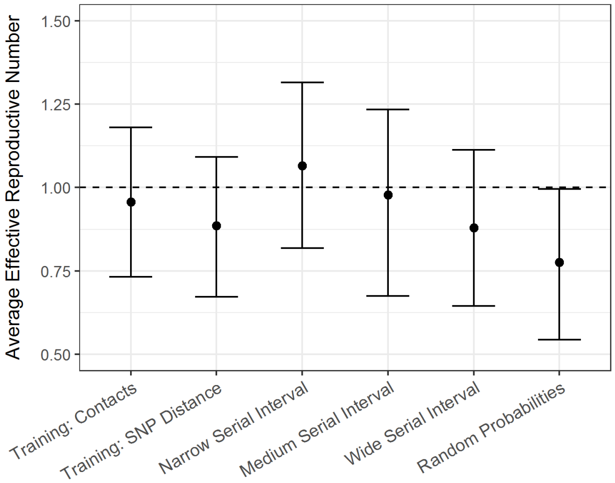

Figure 13: Average effective reproductive number for the Hamburg TB outbreak calculated using the relative transmission probabilities derived from different methods of calculating probabilities: our method and a training set with links based on contact investigation data; our method and a training set with links based on SNP distance; probabilities derived from narrow: gamma(0.54, 1.9), medium: gamma(1.05, 2.0), and wide: gamma(1.33, 3.0) serial interval distributions; and random probabilities. The vertical bars represent 95% bootstrap confidence intervals. The dotted horizontal line represents a $\bar{R}_t$ value of 1.

References

[1] Pierre-Yves Boelle, Severine Ansart, Anne Cori, and Alain-Jacques Valleron. Transmission parameters of the A / H1N1 (2009) influenza virus pandemic : a review. *Influenza and Other Respiratory Viruses*, 5:306–316, 2011. doi: 10.1111/j.1750-2659.2011.00234.x.

[2] Christophe Fraser, Christl A Donnelly, Simon Cauchemez, William P Hanage, Maria D Van Kerkhove, T Déirdre Hollingsworth, Jamie Griffin, Rebecca F Baggaley, Helen E Jenkins, Emily J Lyons, Thibaut Jombart, Wes R Hinsley, Nicholas C Grassly, Francois Balloux, Azra C Ghani, Neil M Ferguson, Andrew Rambaut, Oliver G Pybus, Hugo Lopez-gatell, Celia M Alpuche-aranda, Ietza Bojorquez Chapela, Ethel Palacios Zavala, Francesco Checchi, Erika Garcia, Stephane Hugonnet, and Cathy Roth. Pandemic Potential of a Strain of Influenz A (H1N1): Early Findings. *Science*, 324(JUNE):1557–1561, 2009.

[3] Steven Riley, Christophe Fraser, Christl A Donnelly, Azra C Ghani, Laith J Abu-raddad, Anthony J Hedley, Gabriel M Leung, Lai-ming Ho, Tai-hing Lam, Pak-yin Leung, Thomas Tsang, William Ho, Koon-hung Lee, Edith M. C. Lau, Ferguson, Neil M, and Roy M Anderson. Transmission Dynamics of the Etiological Agent of SARS in Hong Kong : Impact of Public Health Interventions. *Science*, 300 (June):1961–1967, 2003.

[4] Jacco Wallinga and Peter Teunis. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *American Journal of Epidemiology*, 160(6):509–516, 2004.

[5] G Chowell, N W Hengartner, C Castillo-Chavez, P W Fenimore, and J M Hyman. The basic reproductive number of Ebola and the effects of public health measures : the cases of Congo and Uganda. *Journal of Theoretical Biology*, 300(5627):1961–6, 2003. doi: 10.1016/j.jtbi.2004.03.006.

[6] Laura Forsberg White and M Pagano. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Statistics in Medicine*, 27(16):2999–3016, 2008. doi: 10.1002/sim.3136.A.

[7] Y Ma, C R Horsburgh Jr., Laura F White, and Helen E Jenkins. Quantifying TB transmission: a systematic review of reproductive number and serial interval estimates for tuberculosis. *Epidemiology and Infection*, 146(12):1478–1494, 2018.

[8] Margaretha Annelie Vink, Martinus Christoffel, Jozef Bootsma, and Jacco Wallinga. Systematic Reviews and Meta- and Pooled Analyses Serial Intervals of Respiratory Infectious Diseases : A Systematic Review and Analysis. *American Journal of Epidemiology*, 180(9):865–875, 2014. doi: 10.1093/aje/kwu209.

[9] Paul L. Delamater, Erica J. Street, Timothy F. Leslie, Y. Tony Yang, and Kathryn H. Jacobsen. Complexity of the basic reproduction number (R0). *Emerging Infectious Diseases*, 25(1):1–4, 2019. ISSN 10806059. doi: 10.3201/eid2501.171901.

[10] Andreas Roetzer, Roland Deil, Thomas A Kohl, Christian Ruckbert, Ulrick Nubel, Jochen Blom, Thierry Wirth, Sebastian Jaenicke, Sieglinde Schuback, Sabine Rusch-Gerdes, Philip Supply, Jorn Kalinowski, and Stefan Niemann. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak : A Longitudinal Molecular Epidemiological Study. *PLoS Medicine*, 10(2), 2013. doi: 10.1371/journal.pmed.1001387.

[11] Timothy M Walker, Camilla L C Ip, Ruth H Harrell, Jason T Evans, Georgia Kapatai, Martin J Dedicoat, David W Eyre, Daniel J Wilson, Peter M Hawkey, Derrick W Crook, Julian Parkhill, David Harris, A Sarah Walker, Rory Bowden, Philip Monk, E Grace Smith, and Tim E A Peto. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks : a retrospective observational study. *The Lancet Infectious Diseases*, 13:137–146, 2013. ISSN 1473-3099. doi: 10.1016/S1473-3099(12)70277-3. URL `http://dx.doi.org/10.1016/S1473-3099(12)70277-3`.

[12] Robyn S Lee, Nicolas Radomski, Jean-francois Proulx, Ines Levade, B Jesse Shapiro, Fiona Mcintosh, Hafid Soualhine, Dick Menzies, and Marcel A Behr. Population genomics of Mycobacterium tuberculosis in the Inuit. *PNAS*, 112(44):13609–12614, 2015. doi: 10.1073/pnas.1507071112.

[13] Eleanor M Cottam, Gael Thebaud, Jemma Wadsworth, John Gloster, Leonard Mansley, David J Paton, Donald P King, and Daniel T Haydon. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings of the Royal Society*, 275:887–895, 2008. doi: 10.1098/rspb.2007.1442.

[14] Xavier Didelot, David W Eyre, Madeleine Cule, Camilla L C Ip, M Azim Ansari, David Griffiths, Alison Vaughan, Lily O'Connor, Tanya Golubchik, Elizabeth M Batty, Paolo Piazza, Daniel J Wilson, Rory Bowden, Peter J Donnelly, Kate E Dingle, Mark Wilcox, A Sarah Walker, Derrick W Crook, Tim E A Peto, and Rosalind M Harding. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. *Genome Biology*, 13(12), 2013.

[15] T Jombart, R M Eggo, P J Dodd, and F Balloux. Reconstructing disease outbreaks from genetic data : a graph approach. *Heredity*, 106(2):383–390, 2011. ISSN 0018-067X. doi: 10.1038/hdy.2010.78. URL `http://dx.doi.org/10.1038/hdy.2010.78`.

[16] Thibaut Jombart, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser, and Neil Ferguson. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. *PLoS Computational Biology*, 10(1), 2014. doi: 10.1371/journal.pcbi.1003457.

[17] S Wesley Long, Stephen B Beres, Randall J Olsen, and M Musser. Absence of Patient-to-Patient Intrahospital Transmission of Staphylococcus aureus as Determined by Whole-Genome Sequencing. *mBio*, 5(5):1–10, 2014. doi: 10.1128/mBio.01692-14.Editor.

[18] Marco J Morelli, Gael Thebaud, Joel Chadoef, Donald P King, Daniel T Haydon, and Samuel Soubeyrand. A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. *PLoS Computational Biology*, 8(11), 2012. doi: 10.1371/journal.pcbi.1002768.

[19] Colin J Worby, Marc Lipsitch, and William P Hanage. Shared Genomic Variants : Identification of Transmission Routes Using Pathogen Deep-Sequence Data. *American Journal of Epidemiology*, 186 (10):1209–1216, 2017. doi: 10.1093/aje/kwx182.

[20] Colin J Worby, Philip D O'Neill, Theodore Kypraios, Julie V Robotham, Daniela De Angelis, Edward J P Cartwright, Sharon J Peacock, and Ben S Cooper. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. *Annals of Applied Statistics*, 10(1):395–417, 2016.

[21] R J F Ypma, A M A Bataille, A Stegeman, G Koch, J Wallinga, and W M van Ballegooijen. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society*, 279:444–450, 2012. doi: 10.1098/rspb.2011.0913.

[22] Don Klinkenberg, Jantien A Backer, Xavier Didelot, Caroline Colijn, and Jacco Wallinga. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Computational Biology*, 13(5):1–32, 2017.

[23] Xavier Didelot, Christophe Fraser, Jennifer Gardy, and Caroline Colijn. Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks. *Molecular Biology and Evolution*, 34(4): 997–1007, 2017. doi: 10.1093/molbev/msw275.

[24] Finlay Campbell, Anne Cori, Neil Ferguson, Stephen Baker, and Thibaut Jombart. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data [unpublished]. *PLOS Computational Biology*, 15(3):1–20, 2018.

[25] Finlay Campbell, Camilla Strang, Neil Ferguson, Anne Cori, and Thibaut Jombart. When are pathogen genome sequences informative of transmission events? *PLoS Pathogens*, 14(2):1–21, 2018. ISSN 15537374. doi: 10.1371/journal.ppat.1006885.

[26] Ousmane Faye, Pierre Yves Boëlle, Emmanuel Heleze, Oumar Faye, Cheikh Loucoubar, N'Faly Magassouba, Barré Soropogui, Sakoba Keita, Tata Gakou, El Hadji Ibrahima Bah, Lamine Koivogui, Amadou Alpha Sall, and Simon Cauchemez. Chains of transmission and control of Ebola virus disease

in Conakry, Guinea, in 2014: An observational study. *The Lancet Infectious Diseases*, 15(3):320–326, 2015. ISSN 14744457. doi: 10.1016/S1473-3099(14)71075-8.

[27] Zhuang Shen, Fang Ning, Weigong Zhou, Xiong He, Changying Lin, Daniel P Chin, Zonghan Zhu, and Anne Schuchat. Superspreading SARS Events, Beijin, 2003. *Emerging Infectious Diseases*, 10(2), 2004. ISSN 1080-6040. doi: 10.3201/eid1002.030732.

[28] Benjamin Armbruster and Margaret L Brandeau. Contact tracing to control infectious disease: when is enough. *Health Care Management Science*, 10(4):341–355, 2007.

[29] Istvan Z. Kiss, Darren M. Green, and Rowland R. Kao. Disease contact tracing in random and clustered networks. *Proceedings of the Royal Society B: Biological Sciences*, 272(1570):1407–1414, 2005. ISSN 14712970. doi: 10.1098/rspb.2005.3092.

[30] WHO. Global Tuberculosis Report 2018. Technical report, 2018.

[31] Gill Bell and John Potterat. Partner notification for sexually transmitted infections in the modern world: A practitioner perspective on challenges and opportunities. *Sexually Transmitted Infections*, 87 (SUPPL. 2):34–36, 2011. ISSN 13684973. doi: 10.1136/sextrans-2011-050229.

[32] Jonathan E Golub, Wendy A Cronin, Olugbenga O. Obasanjo, William Coggin, Kristina Moore, Diana S Pope, Deidre Thompson, Timothy R Sterling, Susan Harrington, William R Bishai, and Richard E Chaisson. Transmission of Mycobacterium tuberculosis Through Casual Contact With an Infectious Case. *Archives of internal medicine*, 161:2254–2258, 2001.

[33] Roland Diel, Stefan Niemann, and Albert Nienhaus. Risk of tuberculosis transmission among healthcare workers. *ERJ Open Res*, 4(2), 2018. doi: 10.1183/23120541.00161-2017. URL `http://dx.doi.org/10.1183/23120541.00161-2017`.

[34] Roland Diel, Steffen Schneider, Karen Meywald-Walter, Christa-Maria Ruf, Sabine Rüsch-Gerdes, and Stefan Niemann. Epidemiology of Tuberculosis in Hamburg , Germany : Long-Term Population-Based Analysis Applying Classical and Molecular Epidemiological Techniques. *Journal of Clinical Microbiology*, 40(2):532–539, 2002. doi: 10.1128/JCM.40.2.532.

[35] Liangxiao Jiang, Chaoqun Li, Shasha Wang, and Lungan Zhang. Engineering Applications of Artificial Intelligence Deep feature weighting for naive Bayes and its application to text classification. *Engineering Applications of Artificial Intelligence*, 52:26–39, 2016. ISSN 0952-1976. doi: 10.1016/j.engappai.2016.02.002. URL `http://dx.doi.org/10.1016/j.engappai.2016.02.002`.

[36] Ömer Faruk Arar and Kürsat Ayan. A feature dependent Naive Bayes approach and its application to the software defect prediction problem. *Applied Soft Computing*, 59:197–209, 2017. doi: 10.1016/j.asoc.2017.05.043.

[37] Paola Sebastiani, Nadia Solovieff, and Jenny X Sun. Naïve Bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait : not so different after all! *Frontiers in Genetics*, 3:1–9, 2012. doi: 10.3389/fgene.2012.00026.

[38] Klaus Peter Schliep. phangorn: Phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011. ISSN 13674803. doi: 10.1093/bioinformatics/btq706.

[39] James Stimson, Jennifer Gardy, Barun Mathema, Valeriu Crudu, Ted Cohen, and Caroline Colijn. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. *Molecular Biology and Evolution*, 36(3):587–603, 2019.

[40] Emilia Vynnycky and Paul E M Fine. Lifetime Risks, Incubation Period, and Serial Interval of Tuberculosis. *American Journal of Epidemiology*, 152(3):247–263, 2000.

[41] Anne Cori, Pierre Nouvellet, Tini Garske, Herve Bourhy, Emmanuel Nakoune, and Thibaut Jombart. A graph-based evidence synthesis approach to detecting outbreak clusters : An application to dog rabies. *PLoS Computational Biology*, 14(12):1–22, 2018.

[42] Timothy M Walker, Maeve K Lalor, Agnieszka Broda, Luisa Saldana Ortega, Marcus Morgan, Lynne Parker, Sheila Churchill, Karen Bennett, Tanya Golubchik, Adam P Giess, Carlos Del, Ojo Elias, Katie J Jeff, Ian C J W Bowler, Ian F Laurenson, Anne Barrett, Francis Drobniewski, Noel D Mccarthy, Laura F Anderson, Ibrahim Abubakar, H Lucy Thomas, Philip Monk, E Grace Smith, A Sarah Walker, Derrick W Crook, Tim E A Peto, and Christopher P Conlon. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire , UK , 2007–12 , with whole pathogen genome sequences: an observational study. *Lancet Repiratory Medicine*, 2:285–292, 2014. doi: 10.1016/S2213-2600(14)70027-X.

[43] Laura F Anderson, Surinder Tamne, Timothy Brown, John P Watson, Catherine Mullarkey, Dominik Zenner, and Ibrahim Abubakar. Transmission of multidrug-resistant tuberculosis in the UK : a cross-sectional molecular and epidemiological study of clustering and contact tracing. *The Lancet Infectious Diseases*, 14(5):406–415, 2014. ISSN 1473-3099. doi: 10.1016/S1473-3099(14)70022-2. URL `http://dx.doi.org/10.1016/S1473-3099(14)70022-2`.

[44] Anne Marie France, Juliana Grant, J Steve Kammerer, and Thomas R Navin. A Field-Validated Approach Using Surveillance and Genotyping Data to Estimate Tuberculosis Attributable to Recent Transmission in the United States. *American Journal of Epidemiology*, 182(9):799–807, 2015. doi: 10.1093/aje/kwv121.

[45] Josephine M Bryant, Anita C Schürch, Henk Van Deutekom, Simon R Harris, Jessica L De Beer, Victor De Jager, Kristin Kremer, Sacha A F T Van Hijum, Roland J Siezen, Martien Borgdorff, Stephen D Bentley, Julian Parkhill, and Dick Van Soolingen. Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data. *BMC infectious Diseases*, 13(110):1–12, 2013. doi: 10.1186/1471-2334-13-110.

[46] Martien W Borgdorff, Maruschka Sebek, Ronald B Geskus, Kristin Kremer, Nico Kalisvaart, and Dick van Soolingen. The incubation period distribution of tuberculosis estimated with a molecular epidemiological approach. *International Journal of Epidemiology*, 40:964–970, 2011. doi: 10.1093/ije/dyr058.

[47] A H A ten Asbroek, M W Borgdorff, N J D Nagelkerke, M M G G Sebek, W Deville, J D A van Embden, and D van Soolingen. Estimation of serial interval and incubation period of tuberculosis using DNA fingerprinting. *International Journal of Tuberculosis and Lung Disease*, 3(5):414–420, 1999.

[48] Ellen Brooks-Pollock, Mercedes C Becerra, Edward Goldstein, Ted Cohen, and Megan B Murray. Epidemiologic Inference From the Distribution of Tuberculosis Cases in Households in Lima , Peru. *The Journal of Infectious Diseases*, 203:1582–1589, 2011. doi: 10.1093/infdis/jir162.

[49] Christl A Donnelly, Lyn Finelli, Simon Cauchemez, Sonja J Olsen, Saumil Doshi, Michael L Jackson, Erin D Kennedy, Laurie Kamimoto, Tiffany L Marchbanks, Oliver W Morgan, Minal Patel, David L Swerdlow, and Neil M Ferguson. Serial Intervals and the Temporal Distribution of Secondary Infections within Households of 2009 Pandemic Influenza A (H1N1): Implications for Influenza Control Recommendations. *Clinical Infectious Diseases*, 52(Suppl 1):123–130, 2011. doi: 10.1093/cid/ciq028.

[50] Inaki Comas, Susanne Homolka, Stefan Niemann, and Sebastien Gagneux. Genotyping of Genetically Monomorphic Bacteria : DNA Sequencing in Mycobacterium tuberculosis Highlights the Limitations of Current Methodologies. *PLoS ONE*, 4(11), 2009. doi: 10.1371/journal.pone.0007815.

[51] Peter Teunis, Janneke C M Heijne, Faizel Sukhrie, Jan van Eijkeren, Marion Koopmans, and Mirjam Kretzschmar. Infectious disease transmission as a forensic problem : who infected whom? *Journal of the Royal Society Interface*, 10, 2013.

[52] Nesma Settouti, Mohammed El Amine Bechar, and Mohammed Amine Chikh. Statistical Comparisons of the Top 10 Algorithms in Data Mining for Classification Task. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4(1):46–51, 2016. doi: 10.9781/ijimai.2016.419.

[53] Burak Turhan and Ayse Bener. Analysis of Naive Bayes' assumptions on software fault data: An empirical study. *Data & Knowledge Engineering*, 68:278–290, 2009. ISSN 0169-023X. doi: 10.1016/j.datak.2008.10.005. URL `http://dx.doi.org/10.1016/j.datak.2008.10.005`.

[54] Ludmila I Kuncheva. On the optimality of Naive Bayes with dependent binary features. *Pattern Recognition Letters*, 27:830–837, 2006. doi: 10.1016/j.patrec.2005.12.001.

[55] I Rish. An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, pages 41–46, New York, 2001. IBM.

[56] Harry Zhang. The Optimality of Naive Bayes. In *FLAIRS Conference*. AAAI Press, 2004.

[57] Liangxiao Jiang, Dianhong Wang, Zhihua Cai, and Xuesong Yan. Survey of Improving Naive Bayes for Classification. In *International Conference on Advanced Data Mining and Applications*, pages 134–145, Berlin, Heidelberg, 2007. Springer.

[58] Nayyar A Zaidi, Jesus Cerquides, Mark J Carman, and Geoffrey I Webb. Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting. *Journal of Machine Learning Research*, 14:1947–1988, 2013.

[59] Laura F White, Brett Archer, and Marcello Pagano. Determining the dynamics of influenza transmission by age. *Emerging Themes in Epidemiology*, 11(4):1–10, 2014. ISSN Emerging Themes in Epidemiology. doi: 10.1186/1742-7622-11-4. URL `Emerging Themes in Epidemiology`.

[60] Courtney M Yuen, J Steve Kammerer, Kala Marks, Thomas R Navin, and Anne Marie France. Recent Transmission of Tuberculosis — United States , 2011–2014. *PLoS ONE*, 11(4):2011–2014, 2016. doi: 10.1371/journal.pone.0153728.

[61] Chandrashekhar T. Sreeramareddy, Kishore V. Panduru, Joris Menten, and J. Van den Ende. Time delays in diagnosis of pulmonary tuberculosis: A systematic review of literature. *BMC Infectious Diseases*, 9:1–10, 2009. ISSN 14712334. doi: 10.1186/1471-2334-9-91.

[62] Dag Gundersen Storla, Solomon Yimer, and Gunnar Aksel Bjune. A systematic review of delay in the diagnosis and treatment of tuberculosis. *BMC Public Health*, 8(15):1–9, 2008. ISSN 14712458. doi: 10.1186/1471-2458-8-15.