Estimating the Relative Probability of Direct Transmission between Infectious Disease Patients

Sarah V Leavitt *[1], Robyn S Lee[2], Paola Sebastiani[1], Charles R. Horsburgh[3], Helen E Jenkins[1†], 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

† These authors contributed equally

* Corresponding author (sv1205@bu.edu)

ABSTRACT

Background

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.

Methods

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.

Results

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 unlinked pairs with high accuracy (area under the receiver operating curve value of 95%). Additionally only a subset of the cases, 10-50% 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.

Conclusions

Our method is a novel way to infer transmission dynamics in any dataset when only a subset of cases has rich contact investigation and/or genetic data.

Key words: tuberculosis, reproductive number, naive Bayes, machine learning

KEY MESSAGES

- This method provides a way to calculate the relative probability that two infectious disease patients are connected by direct transmission using clinical, demographic, geographic, and genetic characteristics.
- We use a naïve Bayes, a machine learning technique to estimate these probabilities using a training set of probable links defined by contact investigation or pathogen WGS data on a subset of cases.
- These probabilities can be used to explore possible transmission chains, rule out transmission events, and estimate the reproductive number.

INTRODUCTION

Infectious disease parameters such as the serial interval (time between symptom onset from primary to secondary case) and the reproductive number (average number of secondary cases produced by a primary case over the infection course) are instrumental in managing outbreaks (1). For diseases in which disease progression shortly follows infection, these parameters have been studied extensively (1–6). For others, such as tuberculosis (TB), the serial interval and reproductive number estimates are few and inconsistent (7–9).

Methods to estimate both the serial interval and reproductive number often rely on determining which cases are linked by direct transmission. Pathogen whole genome sequence (WGS) data is a powerful tool to link cases and several methods have been developed to analyze these data (10–21). However, WGS data is still relatively expensive and requires bioinformatics expertise, making universal use in high disease burden settings unfeasible and many datasets are likely to have WGS data on only a proportion of cases. Another way to link cases is contact investigations, which are often part of the infectious disease outbreak response (22–28). However, these often do not perfectly identify due to nonspecific transmission mechanisms, disease characteristics, and the willingness and ability of cases to share information about contacts (25,28–32). In addition, contact investigations are time consuming and require significant human resources, again meaning that this data source is unlikely to be available for all cases.

Here, we present a novel 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 proportion of cases, paired with other risk factor data. These probabilities can be used to understand outbreak transmission dynamics and estimate the reproductive number without a reliable serial interval estimate. We apply our method to a TB outbreak in Hamburg, Germany.

METHODS

Data Structure

Our method requires individual-level case data, e.g. geographic location, clinical information, demographics, and sampling date. At least a subset of the cases needs additional information, e.g. detailed contact investigation and/or pathogen genome WGS data, to form the training set to generate the model. We 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 $(\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{p})$ into pair-level covariates $(Z_1, Z_2, \ldots, Z_p)$ by computing "distances" which capture how well the individuals match on covariates values. For example, if the individual-level covariate X_1 was town of residence, the pair-level covariate Z_1 could indicate if the individuals live in the same town, neighboring towns, or more distant towns. This process is described further in the supplementary material.

Naïve Bayes

To estimate the probability that cases i and j are linked by direct transmission, $p(i \rightarrow j)$, we use a classification technique called naive Bayes. This method uses Bayes rule to estimate the probability of an outcome given a set of covariates from the observed frequencies in a training dataset. Our outcome, L_{ij} equals 1 if case i infected case j and 0 otherwise. We know the

probable value of L_{ij} for case-pairs in the training set based on pathogen WGS or contact investigation data and want to predict the probability that $L_{ij} = 1$ for the remaining pairs.

We first use the training set to calculate $P(Z_k = z_k | L = l)$, the probability that the pair-level covariate Z_k equals z_k for each covariate $k \in \{1, 2, \dots, p\}$ for a pair with link status $l \in \{1, 0\}$, using:

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

The indicator function, $\mathbb{1}$, equals 1 if the input is true and 0 if 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 how often a pair i, j has linked status l and value z_k for covariate Z_k, $k \in \{1, 2, \dots, p\}$. Because sparse data can result in zero probabilities, we add 1 to each cell (33). Then we use the training set to calculate $P(L = l)$, the prior probability of link status for $l \in \{1, 0\}$ using:

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

We 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:

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

We calculate the conditional probability of all covariate values given link status $\mathrm{P(Z_1 = z_1, \dots, Z_p = z_p | L = 1)}$, as the product of the conditional probabilities of each covariate, $\mathrm{P(Z_k =}$

$z_k|L = 1)$ for $k \in \{1, 2, \ldots, p\}$, assuming that covariates are conditionally independent (i.e. that the covariate values are independent given link status).

Finally, the estimated probabilities are scaled to represent the relative likelihood that case j has been infected by case i rather than any other sampled case using:

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

We call this scaled probability, $p(i \to j)^s$, the "relative transmission probability". Note: the ordered nature of the pair dataset implies that if case j was observed before case i, then $p(i \to j) = 0$.

Training Dataset Construction

Naive Bayes uses a training set, with a known outcome, to inform a model to estimate probabilities in a separate prediction set. For many infectious diseases we never know the true infector. In our training set, the outcome therefore represents probable rather than certain transmission events, inferred from a subset of cases with pathogen WGS and/or detailed contact investigation data. Because of this uncertainty, we want to estimate the transmission probability of the training case-pairs as well as those which lack WGS or contact data and are only included in the prediction set. Therefore, we use an iterative estimation procedure where each pair has a turn in the prediction set. The algorithm used to create the training dataset and iterative estimation procedure is shown in Figure 1 and described further in the supplementary material.

Reproductive Number Estimation

To estimate the reproductive number, we use the Wallinga and Teunis approach (4) that calculates the relative probability that each case was infected by all other cases using the serial interval distribution. The effective reproductive number (R_i) for each case is then calculated by summing the scaled probabilities for all possible infectees assuming that all cases are sampled and the outbreak is completed with: (7)

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

We use this equation, but with probabilities derived from our naive Bayes approach instead of probabilities based on the serial interval.

By averaging the individual reproductive numbers for all cases observed each month, we obtain monthly effective reproductive number (R_t) estimates and average those values for the stable portion of the outbreak to give an average reproductive number estimate over the study period ($\bar{R}_t$). We calculate confidence intervals for R_t and $\bar{R}_t$ using parametric bootstrapping (see supplementary material).

Simulation Study

We assess our method by applying it to 1000 simulated outbreaks of at least 500 cases, composed of multiple transmission chains using TB transmission dynamics (11,34). We simulate each transmission chain with an $\bar{R}_t$ of 1.2 and a gamma distributed serial interval (shape=1.05, scale=2.0) (Ma, et al, under review). We simulate representative pathogen genomes and inform the model with four different covariates representing clinical and demographic variables and a discretized version of the time between infection.

We compare our method performance when training the model using probable transmission events defined by SNP distances, with training using truly linked and unlinked case-pairs. We also compare the method performance with that of probabilities derived from the time between infection dates and the serial interval distribution motivated by the Wallinga and Teunis method (4). For each simulation scenario (Table 1), we calculate the area under the receiver operating curve (AUC) and assess how the probability of the true infector rank compared to the probabilities of all possible infectors. We also estimate R_t and $\bar{R}_t$. To determine what proportion of cases needs to be in the training set to achieve good performance, we use a sensitivity analysis with various outbreak sizes and training proportions. The simulation structure is explained further in the supplementary material.

Table 1 Here

Hamburg TB Outbreak Application

We apply our method to a small TB outbreak in Hamburg and Schleswig-Holstein, Germany, analyzed in Roetzer et al. (10). The outbreak includes 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 was involved in contact investigations performed by the local health authorities.

We define probable links in the training set in two ways: 1) SNP distances and 2) contact investigation. When training with SNP distance, case-pairs with <2 SNPs are considered linked,

those with >12 SNPs are considered unlinked. Pairs with 2-12 SNPs are excluded from the training set as indeterminate (11,34). When using contact investigation data, pairs that had confirmed contact with each other are considered linked, pairs without confirmed contact with each other are considered unlinked, and cases who did not undergo contact investigation are excluded. For comparison, we also calculate the relative transmission probabilities randomly and using the same serial intervals as the simulation study.

RESULTS

Simulation Study

The sample sizes of the 1000 simulated outbreaks were 500-1178 (median: 545) and each outbreak had 2-39 (median: 14) individual transmission chains with 2-846 (median: 9) cases each. Supplementary Figure S1 shows the relative transmission probability distributions for one outbreak comparing truly linked and unlinked case-pairs. In that outbreak, our method estimated relative transmission probabilities of <0.005 for most unlinked pairs (92% when training with the truth and 89% training using SNP distance). With both ways of defining the training set, our method assigned more than 75% of truly linked case-pairs higher probabilities than the serial interval method (Supplementary Figure S2).

Over 1000 simulations, the average AUC was 97% (standard deviation [SD] 0.6) compared to 95% (SD 1.2) when the model was trained using true links and SNP distances respectively (Figure 2, Supplementary Table S1). When the model was trained with links determined by SNP distances, the estimated probability of the true infector ranked in the top 25% of all possible

source case probabilities 93% (SD 2.4) of the time (compared to 95% (SD 1.6) when training with true links). Our method outperformed probabilities estimated using serial intervals (Figure 2, Supplementary Table S1). Figure 3 and Supplementary Table S2 show the $\bar{R}_t$ estimates for each of the different scenarios compared to the 1.2 value used to simulate the outbreaks. Both our method and the correct serial interval estimated $\bar{R}_t$ accurately. However, when incorrect serial intervals were used, the $\bar{R}_t$ estimates were either too high or too low.

In our sensitivity analysis, the performance improved and the metrics' variability decreased as the proportion of cases in the training set increased (Supplementary Figure S3). If the sample size was at least 500, only 10% of all cases were needed to train the model to obtain good performance. For a sample size of 200-500, training the model with 20% of cases resulted in good performance. For smaller outbreaks, the performance was best with at least 50% of the cases in the training set (Figure 4). The $\bar{R}_t$ estimates grew increasingly accurate as the training dataset proportion increased and when 30% of cases were included, the estimates were close to the true value (Supplementary Figure S4).

Hamburg TB Outbreak Application

Case counts over the course of the outbreak and clinical and demographic characteristics are shown in Figure 5 and Table 2. The 86 cases resulted in 3633 possible ordered case-pairs where the possible infector was observed before the infectee. These pairs were separated by 0-20 SNPs (median=4). Of the 86 individuals, 31 (36%) were part of contact investigations and 51 case-pairs had a confirmed contact. All individual-level covariates were transformed into pair-level covariates for analysis (Table 3).

Tables 2 and 3 here

Figure 6 shows heatmaps of all potential infectors for each infectee using our method compared to random probabilities (Figure 6A) and a serial interval distribution (Figure 6B). Using our method, defining links with either with SNP distance (Figure 6C) or confirmed contact (Figure 6D), there is more variation in the relative transmission probability across possible infectors than the serial interval or random scenarios. Some infectees have infectors with a higher probability than all others in the row, suggesting this is the likely true infector. However, even for rows without a clear single infector, many of the possible infectors have very low probabilities and can be eliminated as the true infector.

All methods except random probabilities show spikes in R_t at the second peak in case counts, but to different degrees (Figure 7). The $\bar{R}_t$ estimate was 0.97 (95% confidence interval [CI] 0.74-1.19) when training with contact investigation and 0.85 (95% CI 0.63-1.06) when training with SNP distances (Supplementary Table S3, Figure 8).

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 unlinked case-pairs in simulation studies. Using a SNP distance proxy for transmission to train the model, the classification accuracy was 95%, and 93% of the time the true infector had a probability in the top 25% of all possible infectors. Our

method outperformed the serial interval method in all metrics and accurately estimated $\bar{R}_t$. This is important because the serial interval is difficult to estimate and highly variable (7,8,35), highlighting the value of estimation methods that are independent of serial interval estimates.

Applying these methods to the Hamburg dataset, we found that both methods allowed for the elimination of many transmission links (Figure 6). The two ways of model training produced slightly different $\bar{R}_t$ estimates, which is expected because neither of the different probable transmission events used to train the model perfectly capture the truth. Using contact investigation for training is more discriminating than SNP distance because we know the cases have actually interacted. However, this may miss links with unknown or unreported contacts. Using SNP distances for training will result in fewer missed links, but could connect cases that never had contact with one another. We hypothesize that the true reproductive number for *M. tuberculosis* in this context lies in between these two estimates (0.84-0.97).

Most established methods for exploring transmission focus on either identifying recent transmission clusters (11,16,34,36–39), recreating possible transmission chains (12,14,15,17–21,40–42), or identifying the true infector (43–47). When estimating transmission parameters, simply knowing clusters is not informative enough and identifying the true infector is often impossible in reality. The strength of our method is that it directly estimates the relative transmission probability for all case-pairs instead of seeking to find the true infector or a set of possible transmission trees. This gives our method broad applicability as it can identify potential true infectors (pairs with very high probabilities) or transmission clusters (groups of pairs with

high probabilities). These probabilities can then be used to estimate transmission parameters incorporating the uncertainty around the true infector.

Other methods to estimate the transmission probabilities either only use genetic data (13) or require prior knowledge of the relationship between the covariates and transmission (48). Our method, however, uses many different information sources about the individual cases without assuming any relationship between these factors and transmission. Additionally, our method uses naive Bayes, a simple but powerful machine learning tool that has many diverse applications (33,49–52). Although traditionally a naive Bayes model is trained with a set of true events, our method performs almost as well when SNP distance is used as a transmission proxy.

Having both a training and prediction set also means that not all cases require highly discriminatory information such as contact investigation or pathogen WGS data to estimate relative transmission probabilities. This is relevant because existing datasets often have rich demographic, clinical, and spatial data but lack detailed contact investigation or pathogen WGS data due to significant time and resources needed to obtain these data. Provided a subset of cases, 10-50% depending on the sample size, has this information, our method can infer transmission patterns among the remaining cases as well.

Our method does make assumptions that might not be appropriate. Firstly, naive Bayes assumes independence of the covariates when conditioning on the outcome, which may not be realistic. However, numerous papers have shown that naive Bayes still performs well even when this assumption is violated (51,53–55). Furthermore, many naive Bayes extensions have been

developed that relax this assumption (33,56,57), which could be easily integrated into our method.

The Wallinga and Teunis (4) approach for estimating $\overline{R_t}$, which we applied, assumes that every case was 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 $\bar{R}_t$estimates, so this is unlikely to be an issue here (4,58). Our probability estimates themselves do not assume all cases in an outbreak are sampled because we estimate the relative probability that one case was infected by another over any other sampled case. However, our method could be affected by biased sampling, e.g. because only certain types of cases are observed or 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, as with other infectious disease analytical approaches, our method assumes that cases were infected in the same order that they were observed (39,59). Although not a strong assumption for diseases with clear symptoms and a short latent period, this may not be appropriate for diseases such as TB, with a highly variable and potentially long latent period, and often substantial delays in care-seeking care and diagnosis (60,61). Although this assumption is a known problem in infectious disease research, it is frequently made (43,44) because relaxing it complicates models substantially.

We have developed a method to estimate the relative transmission probabilities between pairs of cases, which is flexible, using any information sources available without making assumptions

about the relationship between these covariates and transmission. The power of our method is that only a subset of cases requires pathogen WGS or contact investigation data, making this method applicable to many outbreak and surveillance datasets. These probabilities can be used to better understand the transmission dynamics of an outbreak by identifying or ruling out possible transmission events and estimating transmission parameters. In a disease where determining transmission events can be extremely difficult, using transmission probabilities between all possible cases provides a unique and powerful analysis tool.

FUNDING

This work was supported by the US National Institutes of Health [NIHGMS R01GM122876]. SVL 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.

FIGURE LEGENDS

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 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 3: 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 4: 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 5: Case counts by year for the Hamburg outbreak described in Roetzer et al. (10)

Figure 6: 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 7: 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 8: 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.

1. Boelle P-Y, Ansart S, Cori A, Valleron A-J. Transmission parameters of the A / H1N1 (2009) influenza virus pandemic : a review. Influenza Other Respi Viruses. 2011;5:306–16.

2. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. Pandemic Potential of a Strain of Influenz A (H1N1): Early Findings. Science (80-). 2009;324(JUNE):1557–61.

3. Riley S, Fraser C, Donnelly CA, Ghani AC, Abu-raddad LJ, Hedley AJ, et al. Transmission Dynamics of the Etiological Agent of SARS in Hong Kong : Impact of Public Health Interventions. Science (80-). 2003;300(June):1961–7.

4. Wallinga J, Teunis P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. Am J Epidemiol. 2004;160(6):509–16.

5. Chowell G, Hengartner NW, Castillo-Chavez C, Fenimore PW, Hyman JM. The basic reproductive number of Ebola and the effects of public health measures : the cases of Congo and Uganda. J Theor Biol. 2003;300(5627):1961–6.

6. White LF, Pagano M. A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. Stat Med. 2008;27(16):2999–3016.

7. Ma Y, Horsburgh Jr. CR, White LF, Jenkins HE. Quantifying TB transmission: a systematic review of reproductive number and serial interval estimates for tuberculosis. Epidemiol Infect. 2018;146(12):1478–94.

8. Vink MA, Christoffel M, Bootsma J, Wallinga J. Systematic Reviews and Meta- and Pooled Analyses Serial Intervals of Respiratory Infectious Diseases : A Systematic Review and Analysis. Am J Epidemiol. 2014;180(9):865–75.

9. Delamater PL, Street EJ, Leslie TF, Yang YT, Jacobsen KH. Complexity of the basic

reproduction number (R0). Emerg Infect Dis. 2019;25(1):1–4.

10. Roetzer A, Deil R, Kohl TA, Ruckbert C, Nubel U, Blom J, et al. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak : A Longitudinal Molecular Epidemiological Study. PLoS Med. 2013;10(2).

11. Walker TM, Ip CLC, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks : a retrospective observational study. Lancet Infect Dis. 2013;13:137–46.

12. Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. PLoS Comput Biol. 2017;13(5):1–32.

13. Didelot X, Fraser C, Gardy J, Colijn C. Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks. Mol Biol Evol. 2017;34(4):997–1007.

14. Lee RS, Radomski N, Proulx J, Levade I, Shapiro BJ, Mcintosh F, et al. Population genomics of Mycobacterium tuberculosis in the Inuit. PNAS. 2015;112(44):13609–12614.

15. Cottam EM, Thebaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, et al. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. Proc R Soc. 2008;275:887–95.

16. Didelot X, Eyre DW, Cule M, Ip CLC, Ansari MA, Griffiths D, et al. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome Biol. 2013;13(12).

17. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data.

PLoS Comput Biol. 2014;10(1).

18. Long SW, Beres SB, Olsen RJ, Musser M. Absence of Patient-to-Patient Intrahospital Transmission of Staphylococcus aureus as Determined by Whole-Genome Sequencing. MBio. 2014;5(5):1–10.

19. Morelli MJ, Thebaud G, Chadoef J, King DP, Haydon DT, Soubeyrand S. A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. PLoS Comput Biol. 2012;8(11).

20. Worby CJ, O'Neill PD, Kypraios T, Robotham J V, De Angelis D, Cartwright EJP, et al. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. Ann Appl Stat. 2016;10(1):395–417.

21. Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc R Soc. 2012;279:444–50.

22. Faye O, Boëlle PY, Heleze E, Faye O, Loucoubar C, Magassouba N, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: An observational study. Lancet Infect Dis. 2015;15(3):320–6.

23. Shen Z, Ning F, Zhou W, He X, Lin C, Chin DP, et al. Superspreading SARS Events, Beijin, 2003. Emerg Infect Dis. 2004;10(2).

24. Armbruster B, Brandeau ML. Contact tracing to control infectious disease: when is enough. Health Care Manag Sci. 2007;10(4):341–55.

25. Kiss IZ, Green DM, Kao RR. Disease contact tracing in random and clustered networks. Proc R Soc B Biol Sci. 2005;272(1570):1407–14.

26. Campbell F, Cori A, Ferguson N, Baker S, Jombart T. Bayesian inference of transmission

chains using timing of symptoms, pathogen genomes and contact data [unpublished]. PLOS Comput Biol. 2018;15(3):1–20.

27. WHO. Global Tuberculosis Report 2018. 2018.

28. Bell G, Potterat J. Partner notification for sexually transmitted infections in the modern world: A practitioner perspective on challenges and opportunities. Sex Transm Infect. 2011;87(SUPPL. 2):34–6.

29. Golub JE, Cronin WA, Obasanjo OO, Coggin W, Moore K, Pope DS, et al. Transmission of Mycobacterium tuberculosis Through Casual Contact With an Infectious Case. Arch Intern Med. 2001;161:2254–8.

30. Diel R, Niemann S, Nienhaus A. Risk of tuberculosis transmission among healthcare workers. ERJ Open Res. 2018;4(2).

31. Diel R, Schneider S, Meywald-Walter K, Ruf C-M, Rüsch-Gerdes S, Niemann S. Epidemiology of Tuberculosis in Hamburg , Germany : Long-Term Population-Based Analysis Applying Classical and Molecular Epidemiological Techniques. J Clin Microbiol. 2002;40(2):532–9.

32. Oelemann MC, Diel R, Vatin V, Haas W, Rusch-Gerdes S, Locht C, et al. Assessment of an Optimized Mycobacterial Interspersed Repetitive- Unit – Variable-Number Tandem-Repeat Typing System Combined with Spoligotyping for Population-Based Molecular Epidemiology Studies of Tuberculosis. J Clin Microbiol. 2007;45(3):691–7.

33. Jiang L, Li C, Wang S, Zhang L. Engineering Applications of Arti fi cial Intelligence Deep feature weighting for naive Bayes and its application to text classi fi cation. Eng Appl Artif Intell. 2016;52:26–39.

34. Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, et al. Assessment of

Mycobacterium tuberculosis transmission in Oxfordshire , UK , 2007–12 , with whole pathogen genome sequences: an observational study. Lancet Repiratory Med. 2014;2:285–92.

35. Vynnycky E, Fine PEM. Lifetime Risks, Incubation Period, and Serial Interval of Tuberculosis. Am J Epidemiol. 2000;152(3):247–63.

36. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. Mol Biol Evol. 2019;36(3):587–603.

37. Cori A, Nouvellet P, Garske T, Bourhy H, Nakoune E, Jombart T. A graph-based evidence synthesis approach to detecting outbreak clusters : An application to dog rabies. PLoS Comput Biol. 2018;14(12):1–22.

38. Anderson LF, Tamne S, Brown T, Watson JP, Mullarkey C, Zenner D, et al. Transmission of multidrug-resistant tuberculosis in the UK : a cross-sectional molecular and epidemiological study of clustering and contact tracing. Lancet Infect Dis. 2014;14(5):406–15.

39. France AM, Grant J, Kammerer JS, Navin TR. A Field-Validated Approach Using Surveillance and Genotyping Data to Estimate Tuberculosis Attributable to Recent Transmission in the United States. Am J Epidemiol. 2015;182(9):799–807.

40. Jombart T, Eggo RM, Dodd PJ, Balloux F. Reconstructing disease outbreaks from genetic data : a graph approach. Heredity (Edinb). 2011;106(2):383–90.

41. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, von Helden P, et al. Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis : a retrospective. Lancet Respir Med. 2013;1(December):786–92.

42. Worby CJ, Lipsitch M, Hanage WP. Shared Genomic Variants : Identi fi cation of Transmission Routes Using Pathogen Deep-Sequence Data. Am J Epidemiol. 2017;186(10):1209–16.

43. Borgdorff MW, Sebek M, Geskus RB, Kremer K, Kalisvaart N, van Soolingen D. The incubation period distribution of tuberculosis estimated with a molecular epidemiological approach. Int J Epidemiol. 2011;40:964–70.

44. ten Asbroek AHA, Borgdorff MW, Nagelkerke NJD, Sebek MMGG, Deville W, van Embden JDA, et al. Estimation of serial interval and incubation period of tuberculosis using DNA fingerprinting. Int J Tuberc Lung Dis. 1999;3(5):414–20.

45. Brooks-Pollock E, Becerra MC, Goldstein E, Cohen T, Murray MB. Epidemiologic Inference From the Distribution of Tuberculosis Cases in Households in Lima , Peru. J Infect Dis. 2011;203:1582–9.

46. Donnelly CA, Finelli L, Cauchemez S, Olsen SJ, Doshi S, Jackson ML, et al. Serial Intervals and the Temporal Distribution of Secondary Infections within Households of 2009 Pandemic Influenza A (H1N1): Implications for Influenza Control Recommendations. Clin Infect Dis. 2011;52(Suppl 1):123–30.

47. Comas I, Homolka S, Niemann S, Gagneux S. Genotyping of Genetically Monomorphic Bacteria : DNA Sequencing in Mycobacterium tuberculosis Highlights the Limitations of Current Methodologies. PLoS One. 2009;4(11).

48. Teunis P, Heijne JCM, Sukhrie F, van Eijkeren J, Koopmans M, Kretzschmar M. Infectious disease transmission as a forensic problem : who infected whom? J R Soc Interface. 2013;10.

49. Settouti N, Bechar MEA, Chikh MA. Statistical Comparisons of the Top 10 Algorithms in

Data Mining for Classification Task. Int J Interact Multimed Artif Intell. 2016;4(1):46–51.

50. Arar ÖF, Ayan K. A feature dependent Naive Bayes approach and its application to the software defect prediction problem. Appl Soft Comput. 2017;59:197–209.

51. Turhan B, Bener A. Analysis of Naive Bayes' assumptions on software fault data: An empirical study. Data Knowl Eng. 2009;68:278–90.

52. Sebastiani P, Solovieff N, Sun JX. Naïve Bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait : not so different after all! Front Genet. 2012;3:1–9.

53. Kuncheva LI. On the optimality of Naive Bayes with dependent binary features. Pattern Recognit Lett. 2006;27:830–7.

54. Rish I. An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence. New York: IBM; 2001. p. 41–6.

55. Zhang H. The Optimality of Naive Bayes. In: FLAIRS Conference. AAAI Press; 2004.

56. Jiang L, Wang D, Cai Z, Yan X. Survey of Improving Naive Bayes for Classification. In: International Conference on Advanced Data Mining and Applications. Berlin, Heidelberg: Springer; 2007. p. 134–45.

57. Zaidi NA, Cerquides J, Carman MJ, Webb GI. Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting. J Mach Learn Res. 2013;14:1947–88.

58. White LF, Archer B, Pagano M. Determining the dynamics of influenza transmission by age. Emerg Themes Epidemiol. 2014;11(4):1–10.

59. Yuen CM, Kammerer JS, Marks K, Navin TR, France AM. Recent Transmission of Tuberculosis — United States , 2011–2014. PLoS One. 2016;11(4):2011–4.

60. Sreeramareddy CT, Panduru K V., Menten J, Van den Ende J. Time delays in diagnosis of

pulmonary tuberculosis: A systematic review of literature. BMC Infect Dis. 2009;9:1–10.

61. Storla DG, Yimer S, Bjune GA. A systematic review of delay in the diagnosis and treatment of tuberculosis. BMC Public Health. 2008;8(15):1–9.

Figure 1

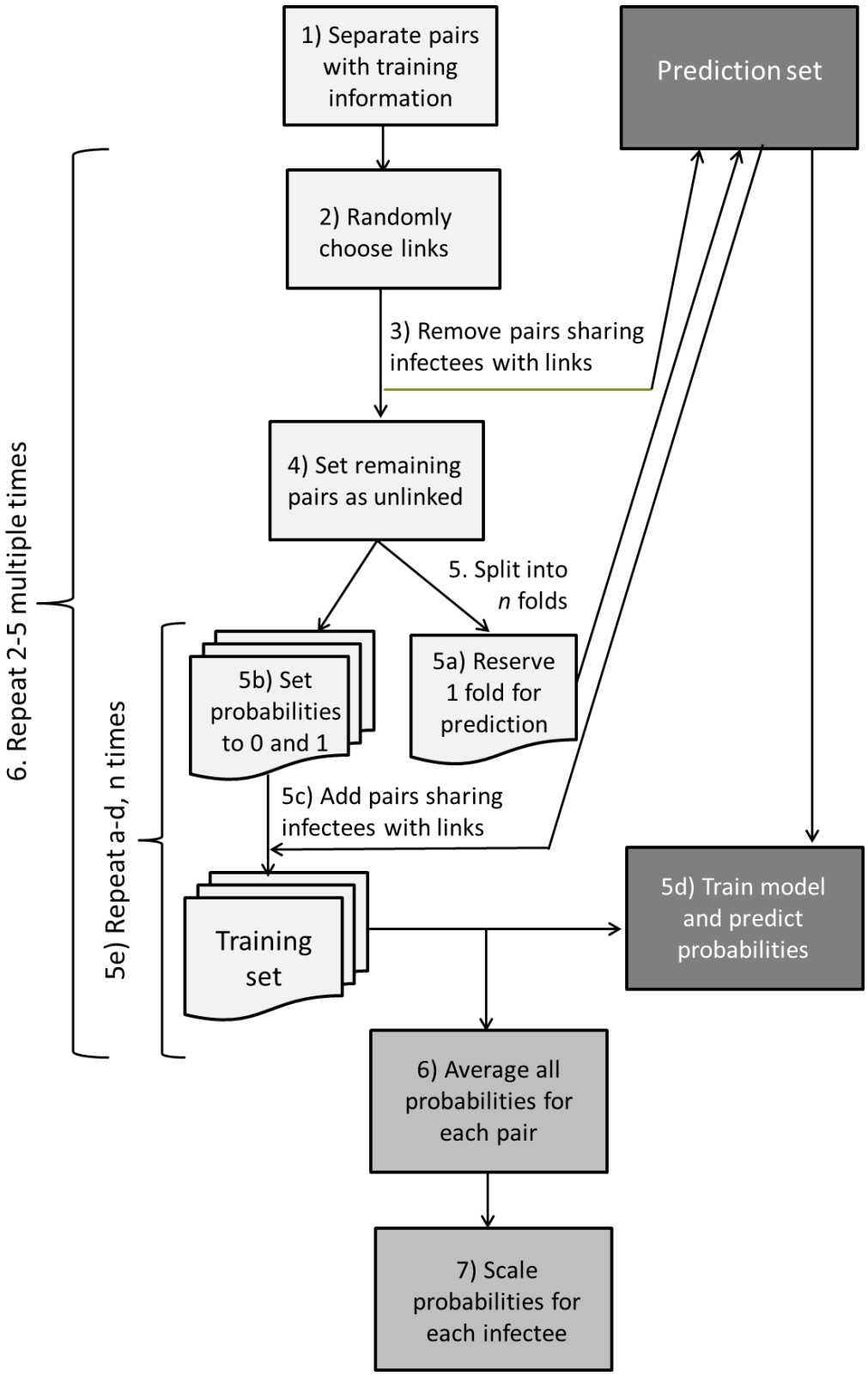

1) Separate pairs with training information
Prediction set
2) Randomly choose links
3) Remove pairs sharing infectees with links
4) Set remaining pairs as unlinked
5. Split into *n* folds
5b) Set probabilities to 0 and 1
5a) Reserve 1 fold for prediction
5c) Add pairs sharing infectees with links
Training set
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 2

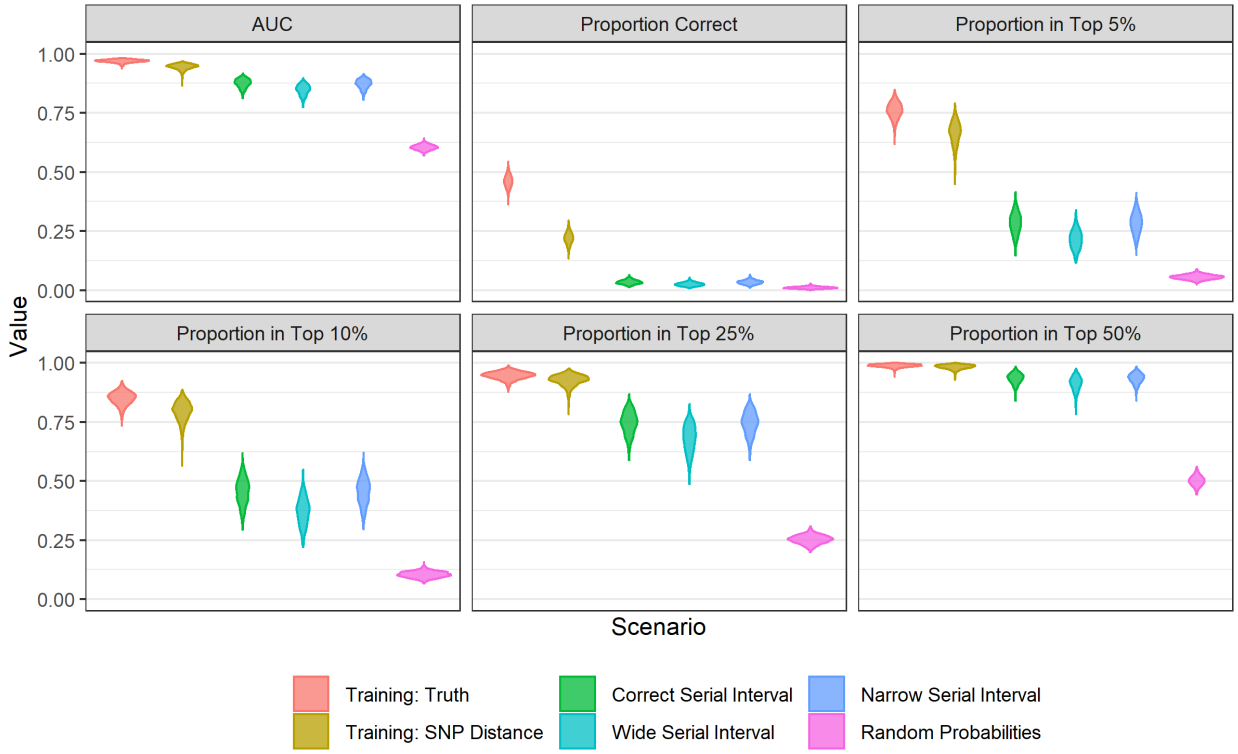

AUC
Proportion Correct
Proportion in Top 5%
Proportion in Top 10%
Proportion in Top 25%
Proportion in Top 50%
Value
1.00
0.75
0.50
0.25
0.00
Scenario
Training: Truth
Training: SNP Distance
Correct Serial Interval
Wide Serial Interval
Narrow Serial Interval
Random Probabilities

Figure 3

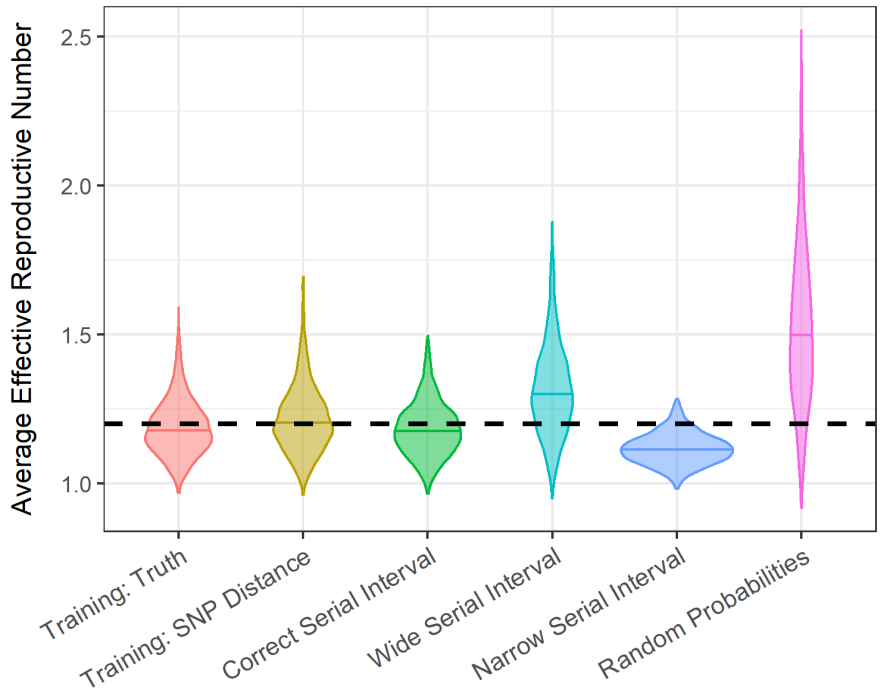


Figure 4

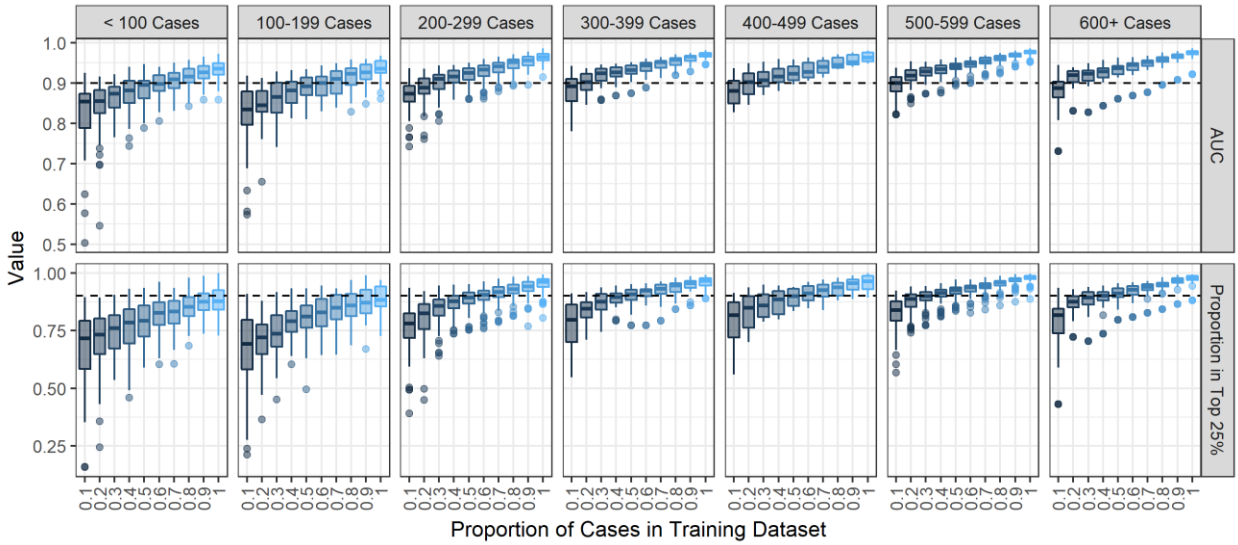


Figure 5

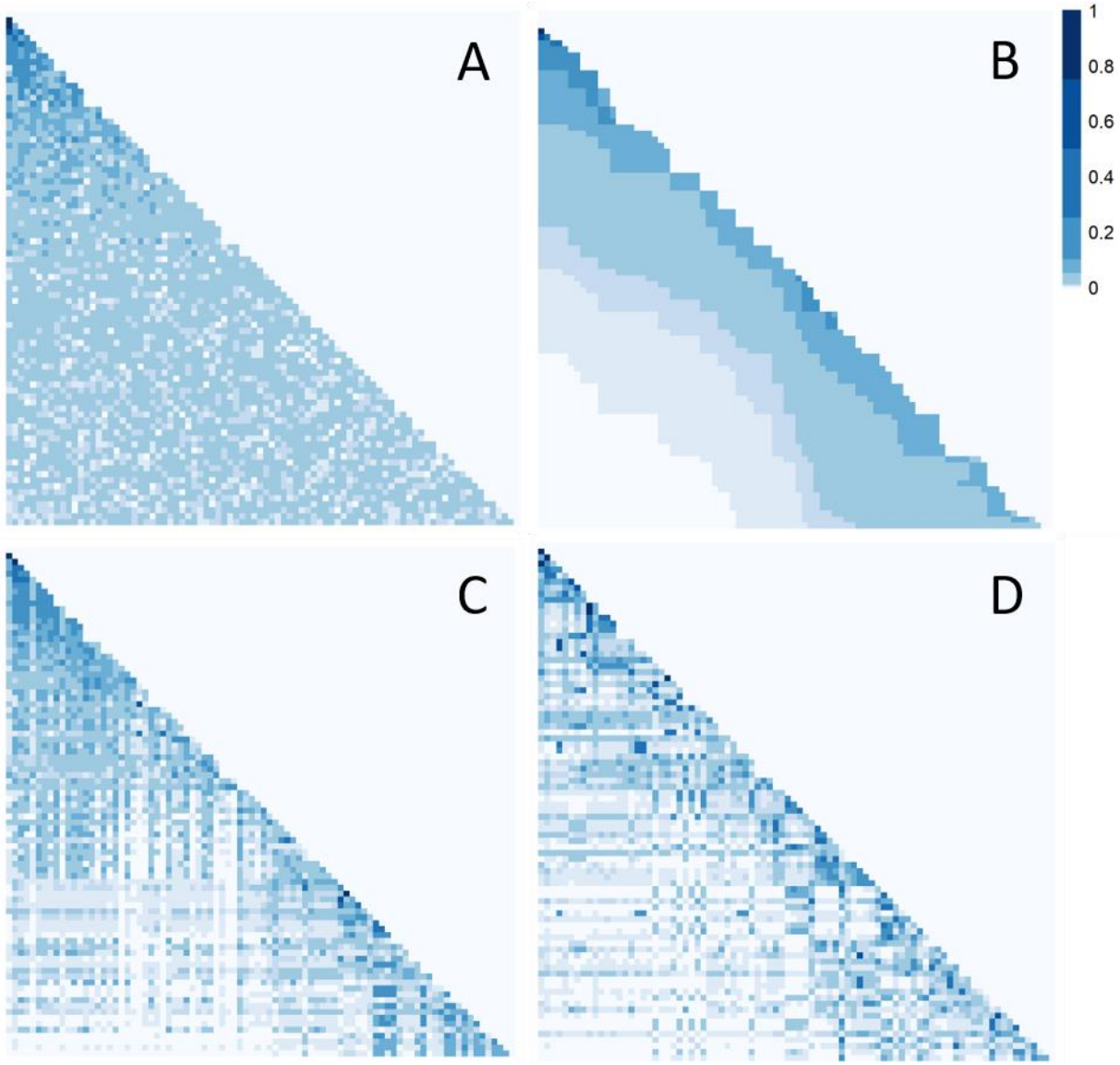


Figure 6

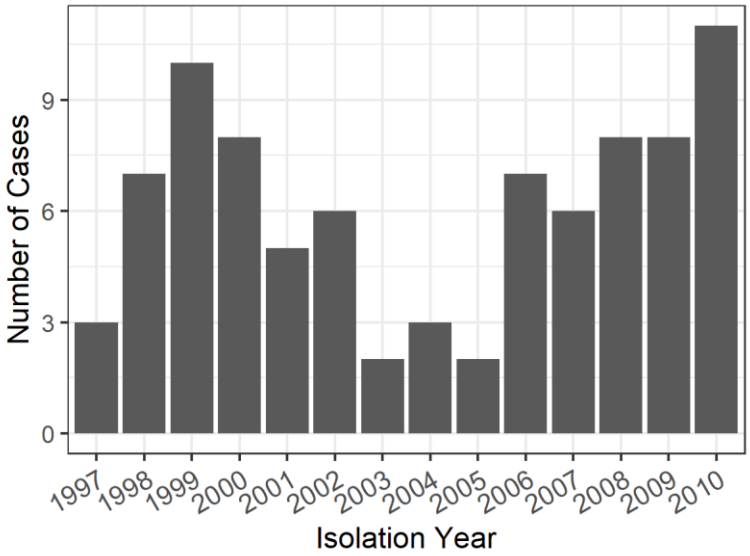


Figure 7

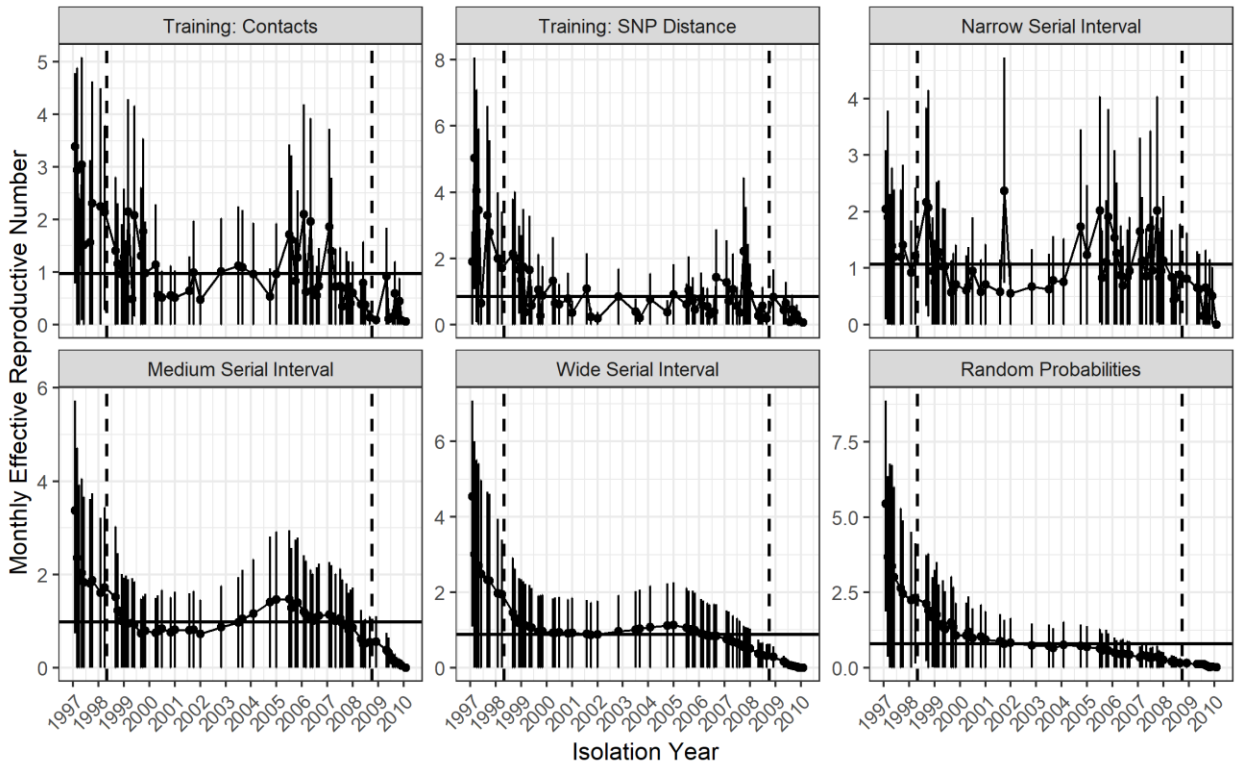
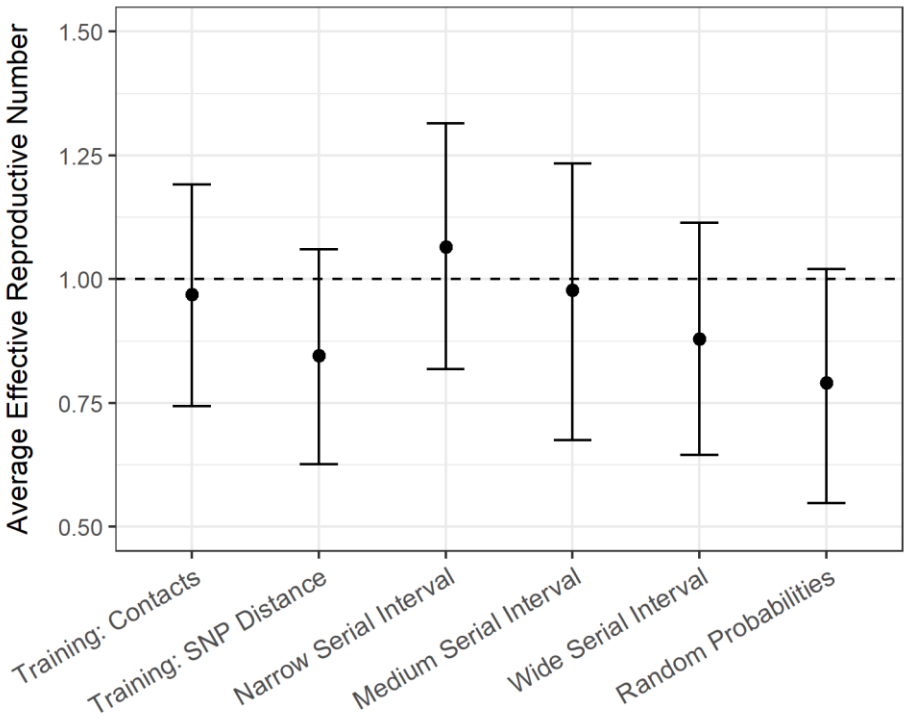


Figure 8


Average Effective Reproductive Number
1.50
1.25
1.00
0.75
0.50
Training: Contacts
Training: SNP Distance
Narrow Serial Interval
Medium Serial Interval
Wide Serial Interval
Random Probabilities