Published in final edited form as:

Epidemiology. 2022 March 01; 33(2): 217–227. doi:10.1097/EDE.0000000000001452.

A Cluster-based Method to Quantify Individual Heterogeneity in Tuberculosis Transmission

Jonathan P. Smith[a,b], **Neel R. Gandhi**[a], **Benjamin J. Silk**[c], **Ted Cohen**[b], **Benjamin Lopman**[a], **Kala Raz**[c], **Kathryn Winglee**[c], **Steve Kammerer**[c], **David Benkeser**[a], **Michael R. Kramer**[a], **Andrew N. Hill**[c]

[a]Emory University Rollins School of Public Health, Atlanta, GA;

[b]Yale University School of Public Health, New Haven, CT;

[c]United States Centers for Disease Control and Prevention, Atlanta, GA.

Abstract

Background: Recent evidence suggests transmission of *Mycobacterium tuberculosis* (Mtb) may be characterized by extreme individual heterogeneity in secondary cases (i.e., few cases account for the majority of transmission). Such heterogeneity implies outbreaks are rarer but more extensive and has profound implications in infectious disease control. However, discrete person-to-person transmission events in tuberculosis (TB) are often unobserved, precluding our ability to directly quantify individual heterogeneity in TB epidemiology.

Methods: We used a modified negative binomial branching process model to quantify the extent of individual heterogeneity using only observed transmission cluster size distribution data (i.e., the simple sum of all cases in a transmission chain) without knowledge of individual-level transmission events. The negative binomial parameter k quantifies the extent of individual heterogeneity (generally, $k < 1$ indicates extensive heterogeneity, and as $k \to \infty$ transmission becomes more homogenous). We validated the robustness of the inference procedure considering common limitations affecting cluster size data. Finally, we demonstrate the epidemiologic utility of this method by applying it to aggregate US molecular surveillance data from the US Centers for Disease Control and Prevention.

Results: The cluster-based method reliably inferred k using TB transmission cluster data despite a high degree of bias introduced into the model. We found that the TB transmission in the United States was characterized by a high propensity for extensive outbreaks ($\hat{k} = 0.09$; 95% confidence interval = 0.09, 0.10).

Correspondence: Jonathan P. Smith, Yale University School of Public Health, 60 College St, New Haven, CT 06511. jonathan.p.smith@yale.edu.

All code to reproduce calculations and simulations in this analysis, including code for implementation of this method in external datasets, is provided on GitHub (https://github.com/jpsmithuga/nbbpClusterAnalysis). US tuberculosis cluster data can be obtained through the Centers for Disease Control and Prevention Department of Tuberculosis Elimination.

The authors report no conflicts of interest.

The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the US Centers for Disease Control and Prevention or the authors' other affiliated institutions.

Supplemental digital content is available through direct URL citations in the HTML and PDF versions of this article (www.epidem.com).

Conclusions: The proposed method can accurately quantify critical parameters that govern TB transmission using simple, more easily obtainable cluster data to improve our understanding of TB epidemiology.

Keywords

Communicable diseases; Disease outbreaks; *Mycobacterium tuberculosis* ; Statistical models

With more than 10 million new cases and 1.4 million deaths in 2019, tuberculosis (TB) is a major contributor to global morbidity and mortality.[1] While the global TB incidence rate has declined over the past 20 years, the rate of decline has recently decelerated.[1,2] This is particularly true for low-incidence settings, where rates of decline have dramatically plateaued.[1] For instance, despite the lowest recorded TB incidence in the world (2.2 cases per 100,000 population), on its current trajectory, the United States will not achieve its goal of TB elimination by the end of this century.[3,4] An improved understanding of factors driving TB incidence across global surveillance systems is needed to guide population-specific, risk-tailored interventions aimed at reaching TB targets.[5,6]

Incident cases of TB arise either through reactivation of a latent TB infection (LTBI) acquired in the distant past or recent transmission. Recent transmission is distinguished from reactivation of LTBI as it focuses on the proportion of infected individuals who progress to TB disease within a relatively short timeframe after infection (e.g., 0–3 years). While reactivation of LTBI is the dominant driver of TB incidence for most low-incidence settings, there remains compelling potential for extensive outbreaks of recent transmission that can fuel larger epidemics and lead to secondary outbreaks elsewhere.[7–12] Hence, in addition to LTBI interventions, preventing transmission remains a key pillar in global TB control programs.

Growing evidence suggests recent transmission is predominantly a result of extreme individual heterogeneity, wherein a small minority of infectious individuals account for the majority of secondary cases (colloquially, “superspreading”).[13–16] Such extreme variation greatly undermines interventions and is an important consideration in prevention strategies.[17–21] Unfortunately, identifying exactly who infected whom among patients with TB is notoriously challenging due to the marked variability in timing of progression from infection to clinical disease. Hence, major gaps in our understanding of individual heterogeneity and its importance in shaping TB epidemiology remain.[14–16]

Individual heterogeneity in transmission is commonly quantified for many infectious diseases by evaluating over-dispersion (e.g., higher than expected variation) in the distribution of secondary cases.[17,22] These methods are not widely applicable to TB since discrete secondary transmission events are unobserved. Fortunately, recent advances in genotyping techniques have afforded the ability for molecular surveillance systems to accurately approximate TB transmission clusters (e.g., all TB cases in a given chain of recent transmission).

Since individual chains of transmission give rise to the final transmission cluster size, there is an inherent relation between the distribution of secondary cases and the distribution of cluster sizes in a given molecular surveillance system. Here, we evaluate a method that relates these two distributions and quantifies the degree of individual heterogeneity in TB transmission using only TB transmission cluster data. We further demonstrate the robustness of the inference procedure under potential limitations arising in TB surveillance and demonstrate the epidemiological significance of the procedure by applying it to aggregate TB molecular surveillance data from the United States.

METHODS

Statistical Methods

This analysis models underlying TB transmission using a single-type branching process. Branching processes are individual-based discrete stochastic processes that are widely used in biology and epidemiology to study the spread of infectious diseases.[17,23–25] Analysis centers on the probability generating function (pgf) of the “offspring” distribution. The offspring distribution is the probability distribution for the number of secondary cases caused by each individual infectious case, denoted z (e.g., $P(Z = z)$ for $z = 0,1,2,\ldots$). The pgf specifies the probabilities associated with each Z value and is defined as $G_Z(s) = \sum_{z=0}^{\infty} P(Z = z)s^z$.[26]

Following previous studies, we assume the offspring distribution follows a negative binomial distribution with mean R_0 (the basic reproduction number, herein referred to as the more generalizable R) and dispersion parameter k.[17] The dispersion parameter k is commonly used in epidemiology to quantify the degree of individual heterogeneity in transmission. Smaller k values ($k < 1$) correspond to increased heterogeneity in secondary cases and imply outbreaks are rarer but more extensive; increasing values of k (e.g., $k > 1$) correspond to more homogeneous transmission. Importantly, the negative binomial converges to the epidemiologically relevant geometric and Poisson distributions when $k = 1$ or $k \rightarrow \infty$, respectively. The geometric distribution corresponds to the underlying assumption of heterogeneity made in typical differential equation models, while the Poisson distribution implies differences in secondary cases is solely attributed to stochasticity.

The primary focus of this analysis is to infer the negative binomial parameter k using only cluster-level data. We approach this by relating the offspring distribution of individual secondary cases, denoted Z, and the offspring distribution of cluster sizes, denoted Y. This relation is initially intuitive; the probability that a chain originating with a single index case results in a final cluster of size $Y = 1$ is identical to the probability of an individual index case results in no secondary transmission, thus $P(Y = 1) = P(Z = 0)$. Expanding to $Y = 2$, the only valid transmission sequence is that an index case results in a single secondary case, thus, $P(Y = 2) = P(Z = 1)P(Z = 0)$. When $Y = 3$, there are only two valid transmission sequences: either the index case results in two secondary cases or the index case results in a single secondary case, who in turn results in a single tertiary case, thus: $P(Y = 3) = P(Z = 2)P(Z = 0)^2 + P(Z = 1)^2\ P(Z = 0)$.

We extend this relation to any cluster of size Y originating with an arbitrary number of n index cases (detailed methods in eAppendix 2; http://links.lww.com/EDE/B887 and eAppendix 3; http://links.lww.com/EDE/B887). Briefly, recall $G_Z(s)$ generates the probabilities that one individual will infect z secondary cases ($z = 0,1,2,3\ldots$). It follows that $G_Z(s)^y$ specifies all possible ways y cases may result in z secondary cases.[27,28] However, two key constraints exist when considering applications of infectious disease transmission. Briefly, only a certain subset of possible permutations from $G_Z(s)^y$ will result in biologically valid transmission sequences. For a cluster originating with n index cases resulting in a final size of exactly y, the proportion of valid transmission sequences is shown to be n/y.[24] Second, for every cluster of size y initiating with n index cases, there must always be exactly $y - n$ transmission events in the cluster regardless of the sequence of transmission. We accounted for these constraints when using the classical procedure to extract the probability from a generating function,[26] resulting in a final probability distribution for a transmission cluster of size y with n index cases defined as:

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

Where $\Gamma(x) = \int_0^\infty t^{x - 1}e^{-t}dt$. This independent derivation extends alternative derivations from Nishiura et al[29] and Blumberg and Lloyd-Smith[30] for the special cases when $R > 1$ and $n = 1$, respectively, and concurs with these derivations when assumptions are met.

Simulated Data

We simulated data to model underlying TB transmission across a range of specified R and k values; values of $k < 1$ are of primary interest and consistent with a high propensity for extreme heterogeneity in transmission. Our primary model assumes empirical values of $R = 0.50$ and $k = 0.15$.[14,31,32] Transmission “chains” are defined as the exact sequence of underlying transmission events (e.g., transmission trees) originating from a single index case. Transmission chains are considered to originate by the sporadic activation of latent TB infection or by the introduction of an infectious individual into the population (e.g., migration). A transmission “cluster” is defined as the final chain size, including the index case and all cases from subsequent generations. An isolated case with no secondary transmission is considered a cluster of size one.

Unless otherwise stated, results are from 500 simulated surveillance systems, each containing 2,000 transmission chains originating with a single index case. Simulated individual-level data contained the full distribution of individual secondary cases (Z values). Final transmission cluster sizes (Y values) were the sum of cases in each transmission chain, including the index case. Thus, simulated cluster data were a simple vector of integers and obscured all information on individual transmission events.

Limitations in TB Surveillance

We modeled several potential real-world limitations affecting cluster size data in TB surveillance (Figure 1 and eAppendix 5; http://links.lww.com/EDE/B887). First, no surveillance system perfectly captures all cases in the population and only a proportion of all cases will be observed (Figure 1A and B). However, the mechanism in which cases are observed is an important consideration affecting the distribution of cluster sizes. Passive surveillance, or the surveillance system's ability to properly identify, diagnose, and report cases, is often the primary mechanism for case ascertainment. Importantly, once a case of TB is identified, many public health systems trigger active case finding measures to identify otherwise unreported cases (i.e., contact tracing). Case ascertainment by active case finding is likely differential by cluster size since larger clusters are more likely to have at least one case observed to trigger active case finding. Evaluating missing cases as a single proportion obscures this phenomenon. We account for this by via. a two-step process. We first simulate passive surveillance by observing each individual within a given transmission chain with probability p_1 (e.g., $p_1 = 0.75$ indicates 75% of all cases are observed by passive surveillance). After evaluation of p_1, if at least one case in a given chain was observed with passive surveillance, we trigger active case finding for that chain. In these chains, all otherwise unobserved cases (i.e., missed by p_1) have an additional opportunity to be observed with probability p_2. Chains that are wholly unobserved are therefore not subject to active case finding. In addition, the position of the missing case in the chain may alter the distribution of cluster sizes (see Figure 1C). To account for this, chains may be "broken" into multiple pseudo-clusters depending on the position of missing cases after evaluation of p_1 and p_2.

Censoring is inherent to analysis of surveillance data; transmission will be ongoing for some unknown proportion of chains at the time of data collection (Figure 1D). Censoring impacts the tail of the distribution as censoring cannot occur for an isolated case (which is either wholly observed or unobserved). To simulate censoring, each chain of size $Y \geq 2$ was randomly designated as censored with probability p_{cens}. The generation where censoring began was randomly selected using a uniform distribution. All cases in the generation selected for censoring and all subsequent generations were unobserved.

Finally, it is often difficult to unambiguously tease apart multiple transmission chains ("overlapping" chains; Figure 1E). This limitation often arises in TB surveillance when cases from two or more transmission chains are geno-typically indistinguishable and thus multiple transmission chains are combined into one genotypic cluster. This results in a single combined cluster of size y with n index cases (or "subclusters"). To simulate overlapping chains, each individual chain was first independently designated as overlapping with probability p_{cens} (e.g., $p_{cens} = 0.10$ indicates 10% of chains in the surveillance system will overlap). Among the pool of chains designated for overlap, we iteratively drew and merged j chains drawn from a Poisson distribution with $\lambda = 2$ (discarding iterations where $j = 0$ or $j = 1$). This process repeated until all chains designated to overlap were merged. Detailed methods and visuals representing the simulation process for all limitations are provided in eAppendix 5 (http://links.lww.com/EDE/B887).

United States TB Data Sources and Definitions

We examined the epidemiologic relevance of this method by applying the inference procedure to data from the US National Tuberculosis Surveillance System (NTSS), the National Tuberculosis Genotyping Service (NTGS), and surveillance for large outbreaks of TB in the United States, which are sponsored by the US Centers for Disease Control and

Prevention (CDC). These data are from all 50 US states and the District of Columbia. Since 2009 CDC has performed genotyping for culture-confirmed cases using 24-locus mycobacterial interspersed repetitive unit-variable number of tandem repeats (MIRU-VNTR) and spacer oligonucleotide typing (spoligotyping) in combination with collecting clinical, demographic, geospatial, and risk factor data for reported TB cases in the United States. Currently, the CDC uses algorithms that consider genotyping results and temporal and spatial proximity to identify clusters of cases that may represent recent transmission. Within this framework, the CDC further identifies possible large outbreaks of 10 or more genotype-matched cases within a 3-year period and monitors them for up to an additional 2 years. Additionally, CDC performs whole genome sequencing (WGS) and compiles local epidemiologic data for all cases detected in possible large outbreaks. WGS provides increased molecular resolution at the level of single nucleotide polymorphisms (SNPs) and can be used alongside epidemiologic data to more precisely identify cases related by recent transmission.

We approximated transmission clusters using genotype-matched cases bound by space and time. We defined four cluster definitions using two timeframes (a 5-year period from 2012 to 2016 and a nested 3-year subset from 2014 to 2016) and two geographic scales (state and county/county equivalent). We considered all clusters size $Y \geq 10$ as censored; we evaluated this assumption by also considering when all clusters of size $Y > 3$ were censored and when no clusters were censored. Genotyping results were obtained for 95.8% of all culture-confirmed TB cases reported in the United States during the 5-year study timeframe.

Large genotype-matched clusters are prone to containing multiple overlapping transmission clusters.[15] To account for this, possible large outbreaks were examined using higher-resolution whole genome sequencing (WGS) data and local epidemiologic data to identify potential overlapping transmission clusters/subclusters (n value). An epidemiologic link was defined as known or probable contact between two patients during either patient's infectious period. Following standard CDC practice for investigation of recent TB transmission in the United States, cases were considered to be in the same subcluster if isolates were within two SNPs or within five SNPs and the cases were epidemiologically linked.[33,34] Cases for which WGS data were not available were included if they were epidemiologically linked to another case in the subcluster. Cases who did not meet these criteria were considered isolated cases within the larger cluster. We evaluated this assumption by assuming all large clusters had $n = 1$ or all $n = y$, representing the minimum and maximum possible values of n, respectively.

The US CDC reviewed this analysis and determined it did not require approval by an institutional review board since data were collected and analyzed as part of routine public health surveillance and determined not to be human subjects research.

Maximum Likelihood Estimation of Transmission Parameters

Maximum likelihood estimation (MLE) was used to jointly estimate transmission parameters, $\hat{R}$ and $\hat{k}$. Confidence intervals (CIs) were obtained using profile likelihood.[35] We used classical methods for MLE and confidence interval estimation for individual-level data as described elsewhere.[22] For cluster-level data, we considered clusters as either wholly observed or censored (i.e., ongoing at the time of data collection). We accounted for censoring by considering censored clusters to be of at least size *y*.[25] The joint likelihood for *a* extinguished clusters and *b* censored clusters is therefore:

$$L(R,k \mid \overrightarrow{a}, \overrightarrow{b}) = \prod_{y=1}^{\infty} \prod_{n=1}^{y} P(Y = y \mid n)^{a_{y,n}} \prod_{y=1}^{\infty} \prod_{n=1}^{y} P(Y \ge y \mid n)^{b_{y,n}}$$

where $P(Y = y|n)$ is the probability density function as specified above and $P(Y \ge y \mid n) = 1 - \sum_{i=1}^{y-1} P(Y = i \mid n)$.

Approach to Evaluating Cluster-based Inference

We first compared the cluster-based inference procedure with classical methods that use the full distribution of secondary cases (i.e., exact transmission chains) under perfect surveillance. Since individual-level data provide the exact number of secondary cases for each case in the transmission chain, *R* and *k* could be directly quantified using standard MLE methods.[22] Using the same dataset, we summed the total number of cases in each chain to generate the distribution of cluster sizes and used the proposed cluster-based MLE method to infer *R* and *k*.

We then evaluated the direction and magnitude of bias arising from the potential limitations in TB surveillance individually. For imperfect case ascertainment, we evaluated each combination of p_1 and p_2 between 0.1 and 1.0. We evaluated bias due to censoring when 5%, 10%, and 20% of clusters were censored (p_{cens} = 0.05,0.10,0.20) and similarly for overlapping clusters (p_{over} = 0.05,0.10,0.20). For each individual analysis, we assumed no bias from other sources.

Finally, we evaluated inference under combined scenarios. Based on published reports and in consultation with global TB surveillance experts, three primary scenarios were developed representing high-resource, moderate-resource, and low-resource settings, with parameter values reflecting increasing bias as resources decline.[36–39] We calculated partial ranked correlation coefficients (PRCCs) under these empirical parametric assumptions (*R* = 0.50, *k* = 0.15) to evaluate the strength of the relation between each limitation and its effect on *k*.

RESULTS

Initial Validation of the Inference Procedure

Table 1 compares the proposed cluster-based inference method with classical methods that utilize the full distribution of individual secondary cases[22] across a range of *R* and *k* values under perfect surveillance. For all values, MLE estimates for both *R* and *k* from the cluster-based inference procedure were unbiased and sufficient in parameter inference (see also eFigure 1; http://links.lww.com/EDE/B887). Under our empirical model assumptions (true *R* = 0.50, *k* = 0.15), the cluster-based inference procedure accurately estimated $\hat{R} = 0.50$ (95% confidence interval [CI] = 0.45, 0.55) and $\hat{k} = 0.15$ (CI = 0.14, 0.16).

Bias Arising Due to Individual Limitations in Surveillance

Missing cases result in a systematic overestimation of $\hat{k}\,(\mathrm{i.e.}, \hat{k} > k)$, biasing estimates towards homogeneity (Figure 2). Our empirical model only slightly overestimated k despite a high degree of bias introduced into the model: $\hat{k} = 0.18$ (CI = 0.14, 0.23) when 50% of cases were missing (p_1 = 0 50) and ignoring active case finding (p_2 = 0.00). Additional case ascertainment through active case finding (increasing p_2) overinflated $\hat{k}$ (see Figure 2 and eFigure 2A; http://links.lww.com/EDE/B887). Missing cases had minimal impact on R (eFigure 2B; http://links.lww.com/EDE/B887). Censored clusters systematically underestimated $k(\mathrm{i.e.}, \hat{k} < k)$, biasing estimates towards heterogeneity. However, under empirical model assumptions this bias was negligible across all censoring thresholds: $\hat{k} = 0.15$ (CI = 0.14, 0.16), 0.14 (CI = 0.13, 0.16), and 0.13 (CI = 0.12, 0.15) when 5%, 10%, and 20% of clusters were censored, respectively. Accounting for this bias in the joint likelihood demonstrated a modest correction in the inference of k (Figure 3 and eFigure 3; http://links.lww.com/EDE/B887).

Inference of k was sensitive to overlapping clusters (Figure 4 and eFigure 4; http://links.lww.com/EDE/B887). Without accounting for overlapping clusters, $\hat{k}$ was dramatically biased upward. Under empirical model assumptions, $\hat{k} = 0.23$ (CI = 0.20, 0.28), 0.36 (CI = 0.30, 0.45), and 1.46 (CI = 1.0, 2.4) at threshold values of 5%, 10%, and 20% of clusters in the surveillance system overlapping, respectively. Our approach to correcting this bias by conditioning the likelihood on the number of index cases sufficiently corrected for this bias ($\hat{k} = 0.15$ [CI = 0.14, 0.16] for all three thresholds).

Performance of Inference Procedure Under Combined Scenarios

We modeled TB transmission under combined imperfect surveillance using high-resource, moderate-resource, and low-resource definitions (Table 2). Inference of both R and k was robust and could clearly and reliably distinguish between small differences in R and k values across all scenarios (Figure 5 and eFigure 5; http://links.lww.com/EDE/B887). Importantly, all three scenarios could unequivocally distinguish between $k = 1$, representing the geometric distribution, and all values below 0.50, including the empirical estimate of $k = 0.15$. There was a slight overestimation of k across all scenarios, which systematically increases as the true underlying value of k increases. $\hat{R}$ was consistently accurate and robust in all models (Figure 5 and eFigure 5; http://links.lww.com/EDE/B887, x axis).

Passive surveillance had a moderate effect and was most influential on model estimates under empirical assumptions (eFigure 6; http://links.lww.com/EDE/B887; PRCC −0.594, $P < 0.001$). Coverage probabilities were calculated to validate the simulation procedure for each scenario and demonstrated minimal bias (eTable 1; http://links.lww.com/EDE/B887 and eFigure 7; http://links.lww.com/EDE/B887).

Analysis of US Surveillance Data

In the full 5-year timeframe (2012–2016), the United States reported 35,313 genotyped cases of TB. There were 29,238 genotypic clusters when defined at the county level (75% were isolated cases considered “clusters” of size 1), and 26,999 clusters (81% of size

1) when defined at the state level (eTable 2; http://links.lww.com/EDE/B887). The 3-year (2014–2016) subset included 20,780 cases of TB resulting in 18,128 clusters when defined at the county level and 16,212 clusters at the state level. Estimates using our primary definition for transmission cluster (5-year, county level) yielded $\hat{k} = 0.09$ (CI = 0.08, 0.09). Inference of k remained robust throughout all four scenarios, ranging from 0.08 (3-year, county level) to 0.12 (5-year, state level; Table 3 and eFigure 8; http://links.lww.com/EDE/B887).

DISCUSSION

Using both empirical and simulated data, we validated a method to quantify the parameters that govern TB transmission dynamics without the need for information on individual-level transmission events. Using more easily obtainable transmission cluster size data, we found inference of individual heterogeneity remained robust despite potential real-world surveillance limitations. We applied this method to TB genotyping cluster data in the United States using multiple transmission cluster definitions and found values of $\hat{k}$ were consistent with a high propensity for extensive outbreaks.

To our knowledge, this study is the first to quantify individual heterogeneity in TB transmission in the United States, and these results are consistent with estimates of $\hat{k}$ in other low-incidence populations.[13,14] Melsew et al[14] used comprehensive follow-up data in Victoria, Australia, to estimate k directly from detailed individual-level data on secondary cases ($\hat{k} = 0.036$). This study provides strong evidence of transmission heterogeneity, yet such detailed data are rarely available in a surveillance setting. Ypma et al[13] estimated heterogeneity from cluster-level surveillance data in the Netherlands by relating individual variation specifically to the distribution of *IS*6110 restriction fragment length polymorphism (RFLP) genotypic cluster sizes ($\hat{k} = 0.10$). RFLP is less discriminatory than MIRU-VNTR and WGS; as a consequence, the authors note their methods preclude accurate inference of R from the data. This complicates the interpretation of results, as both R and k are needed to accurately describe the propensity for large outbreaks in the population (eFigure 9; http://links.lww.com/EDE/B887). The methods proposed here build upon these foundational studies to more accurately characterize the propensity for a large outbreak.

Our results show that the accuracy of parameter estimation is more likely affected by potential limitations in surveillance than by biased inference, and accurate identification of transmission clusters and index cases is paramount to the utility of these methods. Notably, we show improved case detection through active case finding paradoxically makes transmission appear more homogeneous when using cluster-based inference (biases $\hat{k}$ upward) due to small or isolated clusters being more likely to be wholly unobserved by passive surveillance and subsequently not eligible for active case finding measures. We also found the degree of uncertainty in parameter estimation is an increasing function of k itself; larger underlying k values show broader confidence intervals around $\hat{k}$. This is likely because, as k increases, individual differences in transmission become more attributed to stochasticity rather than the underlying mechanisms of disease transmission.

Our model was a simplified representation of recent TB transmission and assumed transmission is independent and identically distributed. We assumed mean susceptibility between individuals remained constant. In reality highly susceptible individuals generally acquire infection first, and thus average susceptibility reduces over generations of spread.[43] These results should be interpreted cautiously in smaller populations where the depletion of susceptible individuals may impact average susceptibility, which tends to decrease the effective reproduction number R_t.[44,45] NTSS data are not real-time data, and the database is reliant upon state and local jurisdictions accurately diagnosing and reporting TB cases. We modeled molecular surveillance data with transmission clusters defined imprecisely using conventional genotyping. In particular, genotype clustering defined using MIRU-VNTR is less discriminatory for cases within the East Asian Mycobacterium tuberculosis lineage and certain other endemic genotypes that are prevalent in the United States, as evidenced by the increased molecular resolution of WGS and phylogenetic analysis results that can identify smaller clusters and isolated cases.[46] Moreover, sources of heterogeneity were not explicitly considered, and imported incident cases with the same MIRU-VNTR may appear to cluster. Recent work has shown that alternative distributions, such as the Poisson-lognormal, may provide better statistical fits for genomic cluster data.[47] In addition, WGS and epidemiologic link data were imperfectly matched to the NTSS and NTSG cluster distribution data and may overestimate the number of index cases in the original cluster. However, varying the number of index cases in large clusters from $n = 1$ to a maximum of $n = y$ revealed that the estimates of k were largely insensitive to these assumptions.

This analysis provides a well-characterized model using simplified data to quantify individual differences in the number of secondary TB cases. This information has been notably absent from TB epidemiology yet is critical to surveillance systems seeking to better understand the underlying mechanisms of TB transmission. The application of this method also affords the opportunity to develop de novo epidemic models that better account for individual heterogeneity when evaluating prevention measures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Seth Blumberg and Rolf Ypma for their consultation.

This work was supported in part by the National Center for HIV, Viral Hepatitis, STD, and TB Prevention Epidemiologic and Economic Modeling Agreement (NEEMA) from the US Centers for Disease Control and Prevention.

REFERENCES

1. World Health Organization. Global Tuberculosis Report 2020. 2020. Available at: https://www.who.int/publications/i/item/9789240013131. Accessed 1 December 2021.
2. Stop TB Partnership. Global Plan to End TB: The Paradigm Shift 2016–2020. 2015. Available at: https://stoptb.org/assets/documents/global/plan/globalplantoendtb_theparadigmshift_2016-2020_stoptbpartnership.pdf. Accessed 01 December 2021.

3. Deutsch-Feldman M, Pratt RH, Price SF, Tsang CA, Self JL. Tuberculosis - United States, 2020. MMWR Morb Mortal Wkly Rep. 2021;70:409–414. [PubMed: 33764959]
4. Menzies NA, Cohen T, Hill AN, et al. Prospects for tuberculosis elimination in the United States: results of a transmission dynamic model. Am J Epidemiol. 2018;187:2011–2020. [PubMed: 29762657]
5. Mathema B, Andrews JR, Cohen T, et al. Drivers of tuberculosis transmission. J Infect Dis. 2017;216(suppl_6):S644–S653. [PubMed: 29112745]
6. Trauer JM, Dodd PJ, Gomes MGM, et al. The Importance of heterogeneity to the epidemiology of tuberculosis. Clin Infect Dis. 2019;69:159–166. [PubMed: 30383204]
7. Althomsons SP, Kammerer JS, Shang N, Navin TR. Using routinely reported tuberculosis genotyping and surveillance data to predict tuberculosis outbreaks. PLoS One. 2012;7:e48754. [PubMed: 23144956]
8. Cohen T, Colijn C, Finklea B, Murray M. Exogenous re-infection and the dynamics of tuberculosis epidemics: local effects in a network model of transmission. J R Soc Interface. 2007;4:523–531. [PubMed: 17251134]
9. Yuen CM, Kammerer JS, Marks K, Navin TR, France AM. Recent transmission of tuberculosis - United States, 2011–2014. PLoS One. 2016;11:e0153728. [PubMed: 27082644]
10. Connors WJ, Hussen SA, Holland DP, Mohamed O, Andes KL, Goswami ND. Homeless shelter context and tuberculosis illness experiences during a large outbreak in Atlanta, Georgia. Public Health Action. 2017;7:224–230. [PubMed: 29018769]
11. Zmak L, Obrovac M, Lovric Z, Jankovic Makek M, Katalinic Jankovic V. Neglected disease in mentally ill patients: major tuberculosis outbreak in a psychiatric hospital. Am J Infect Control. 2017;45:456–457. [PubMed: 27769707]
12. Norheim G, Seterelv S, Arnesen TM, et al. Tuberculosis outbreak in an educational institution in Norway. J Clin Microbiol. 2017;55:1327–1333. [PubMed: 28202795]
13. Ypma RJ, Altes HK, van Soolingen D, Wallinga J, van Ballegooijen WM. A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. Epidemiology. 2013;24:395–400. [PubMed: 23446314]
14. Melsew YA, Gambhir M, Cheng AC, et al. The role of super-spreading events in Mycobacterium tuberculosis transmission: evidence from contact tracing. BMC Infect Dis. 2019;19:244. [PubMed: 30866840]
15. Gardy JL, Johnston JC, Ho Sui SJ, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 2011;364:730–739. [PubMed: 21345102]
16. McCreesh N, White RG. An explanation for the low proportion of tuberculosis that results from transmission between household and known social contacts. Sci Rep. 2018;8:5382. [PubMed: 29599463]
17. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature. 2005;438:355–359. [PubMed: 16292310]
18. Wong G, Liu W, Liu Y, Zhou B, Bi Y, Gao GF. MERS, SARS, and Ebola: the role of super-spreaders in infectious disease. Cell Host Microbe. 2015;18:398–401. [PubMed: 26468744]
19. Kucharski AJ, Althaus CL. The role of superspreading in Middle East respiratory syndrome coronavirus (MERS-CoV) transmission. Euro Surveill. 2015;20:14–18. [PubMed: 26132768]
20. Dye C, Gay N. Epidemiology. Modeling the SARS epidemic. Science. 2003;300:1884–1885. [PubMed: 12766208]
21. Lipsitch M, Cohen T, Cooper B, et al. Transmission dynamics and control of severe acute respiratory syndrome. Science. 2003;300:1966–1970. [PubMed: 12766207]
22. Lloyd-Smith JO. Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. PLoS One. 2007;2:e180. [PubMed: 17299582]
23. Dubrow R Introduction to Stochastic Processes With R. Wiley; 2016.
24. Becker N On parametric estimation for mortal branching processes. Biometrika. 1974;61:393–399.
25. Farrington CP, Kanaan MN, Gay NJ. Branching process models for surveillance of infectious diseases controlled by mass vaccination. Biostatistics. 2003;4:279–295. [PubMed: 12925522]

26. Yan P Distribution theory, stochastic processes, and infectious disease modeling. In: Brauer F, Driessche vd, Wu J, eds. Mathematical Epidemiology. Springer; 2008:229–293.

27. Harris TE. The Theory of Branching Process. Springer-Verlag; 1963.

28. Lange K Applied Probability. 2nd ed. Springer; 2010.

29. Nishiura H, Yan P, Sleeman CK, Mode CJ. Estimating the transmission potential of supercritical processes based on the final size distribution of minor outbreaks. J Theor Biol. 2012;294:48–55. [PubMed: 22079419]

30. Blumberg S, Lloyd-Smith JO. Inference of R(0) and transmission heterogeneity from the size distribution of stuttering chains. PLoS Comput Biol. 2013;9:e1002993. [PubMed: 23658504]

31. Salpeter EE, Salpeter SR. Mathematical model for the epidemiology of tuberculosis, with estimates of the reproductive number and infection-delay function. Am J Epidemiol. 1998;147:398–406. [PubMed: 9508108]

32. Borgdorff MW, Behr MA, Nagelkerke NJ, Hopewell PC, Small PM. Transmission of tuberculosis in San Francisco and its association with immigration and ethnicity. Int J Tuberc Lung Dis. 2000;4:287–294. [PubMed: 10777075]

33. Talarico S, Silk BJ. Whole Genome Sequencing for Investigation of Recent TB Transmission in the United States: Current Uses and Future Plans. Atlanta, GA: Centers for Disease Control and Prevention (CDC), 2018. Availbale at: https://www.cdc.gov/tb/programs/genotyping/pdf/TuberculosiswgSNP-Training-Slides.pdf. Accessed 1 December 2021.

34. Winglee K, McDaniel CJ, Linde L, et al. Logically inferred tuberculosis transmission (LITT): a data integration algorithm to rank potential source cases. Front Public Health. 2021;9:667337. [PubMed: 34235130]

35. Venzon DJ, Moolgavkar SH. A method for computing profile-likelihood-based confidence intervals. J Royal Stat Soc Series C (Applied Statistics). 1988;37:87–94.

36. Saunders MJ, Tovar MA, Collier D, et al. Active and passive case-finding in tuberculosis-affected households in Peru: a 10-year prospective cohort study. Lancet Infect Dis. 2019;19:519–528. [PubMed: 30910427]

37. Mor Z, Migliori GB, Althomsons SP, Loddenkemper R, Trnka L, Iademarco MF. Comparison of tuberculosis surveillance systems in low-incidence industrialised countries. Eur Respir J. 2008;32:1616–1624. [PubMed: 18684850]

38. Doyle TJ, Glynn MK, Groseclose SL. Completeness of notifiable infectious disease reporting in the United States: an analytical literature review. Am J Epidemiol. 2002;155:866–874. [PubMed: 11978592]

39. Wood R, Middelkoop K, Myer L, et al. Undiagnosed tuberculosis in a community with high HIV prevalence: implications for tuberculosis control. Am J Respir Crit Care Med. 2007;175:87–93. [PubMed: 16973982]

40. Keramarou M, Evans MR. Completeness of infectious disease notification in the United Kingdom: a systematic review. J Infect. 2012;64:555–564. [PubMed: 22414684]

41. Zhou D, Pender M, Jiang W, Mao W, Tang S. Under-reporting of TB cases and associated factors: a case study in China. BMC Public Health. 2019;19:1664. [PubMed: 31829147]

42. Haraka F, Glass TR, Sikalengo G, et al. A bundle of services increased ascertainment of tuberculosis among HIV-infected individuals enrolled in a HIV cohort in rural Sub-Saharan Africa. PLoS One. 2015;10:e0123275. [PubMed: 25897491]

43. Hougaard P Life table methods for heterogeneous populations: distributions describing the heterogeneity. Biometrika. 1984;71:75–83.

44. Becker N, Marschner I. The effect of heterogeneity on the spread of disease. In: Gabriel JP, Lefèvre C, Picard P, eds. Stochastic Processes in Epidemic Theory. Lecture Notes in Biomathematics (vol 86). Springer; 1990. Available at: https://link.springer.com/chapter/10.1007%2F978-3-662-10067-7_9. Accessed 1 December 2021.

45. Karlin S, Taylor HM. A First Course in Stochastic Processes. 2nd ed. Academic Press; 1975.

46. Merker M, Blin C, Mona S, et al. Evolutionary history and global spread of the Mycobacterium tuberculosis Beijing lineage. Nat Genet. 2015;47:242–249. [PubMed: 25599400]

47. Brooks-Pollock E, Danon L, Korthals Altes H, et al. A model of tuberculosis clustering in low incidence countries reveals more transmission in the United Kingdom than the Netherlands between 2010 and 2015. PLoS Comput Biol. 2020;16:e1007687. [PubMed: 32218567]

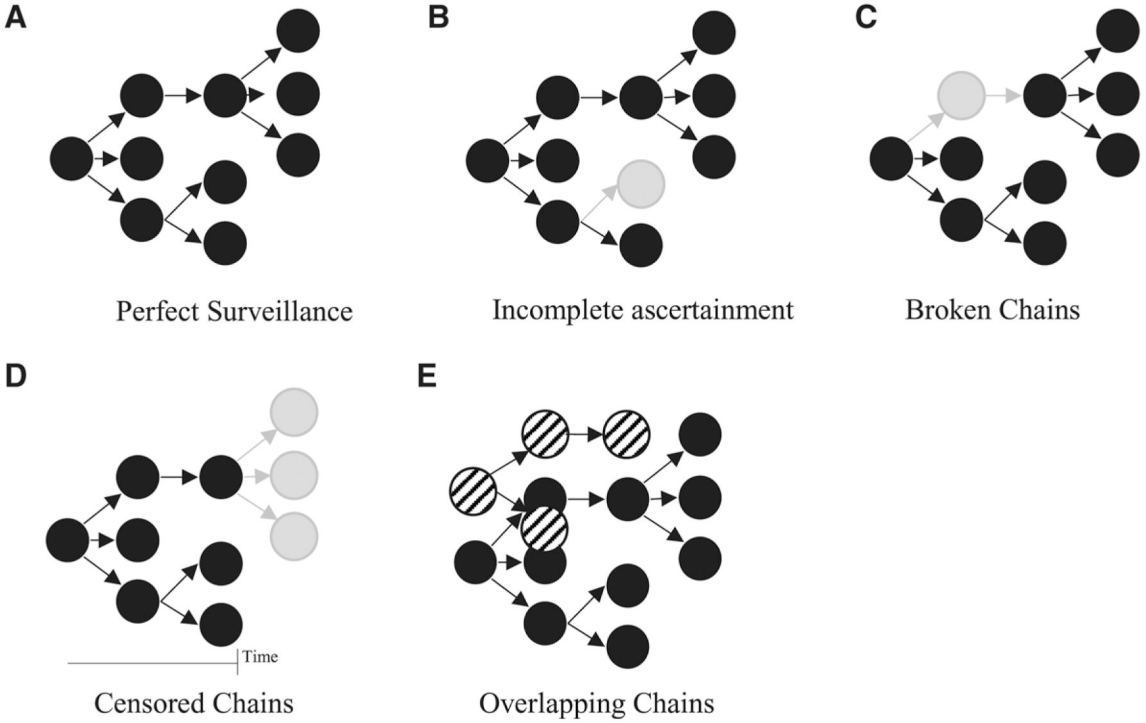
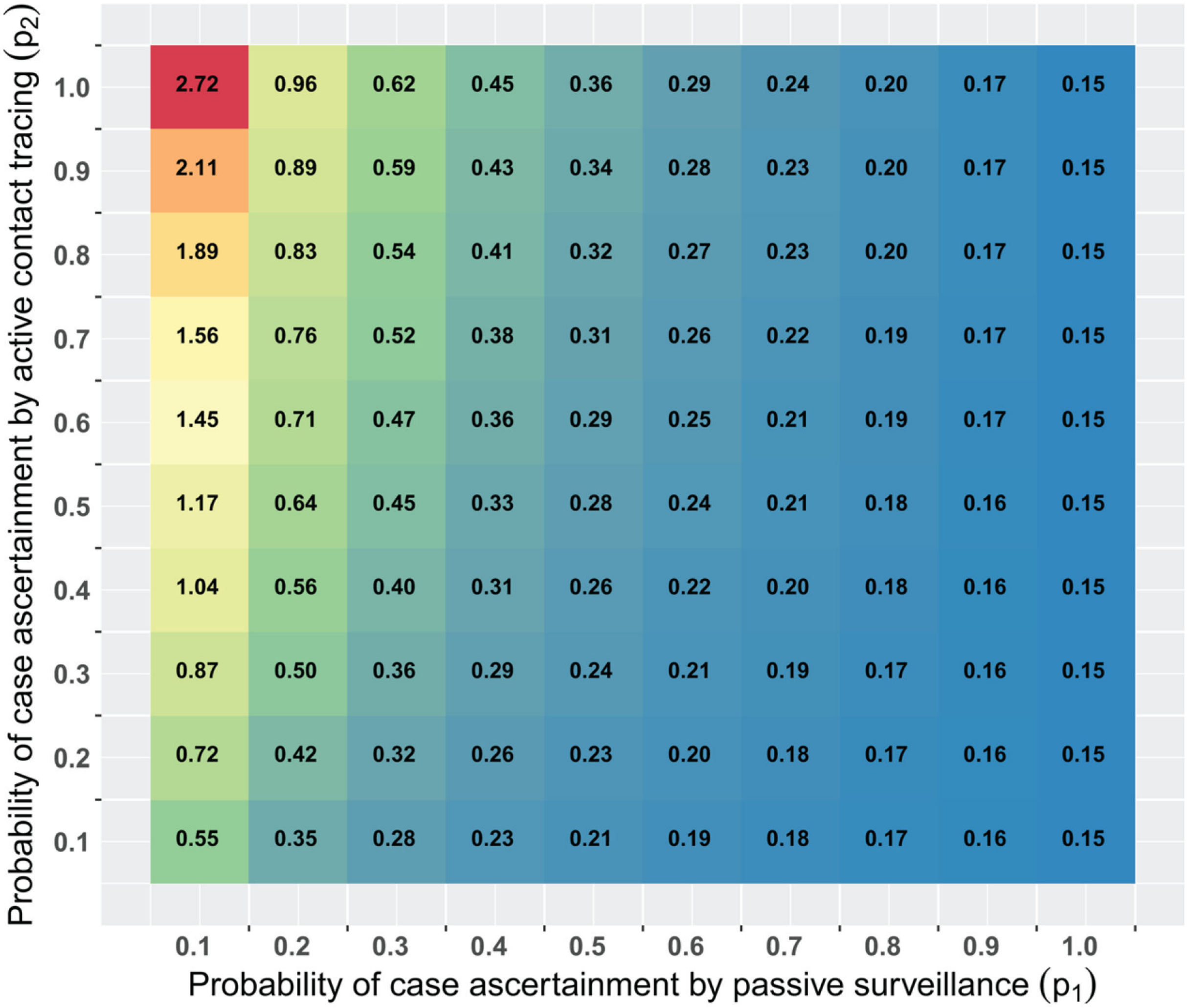


FIGURE 1.

Visualizing limitations arising in tuberculosis transmission surveillance. In the below transmission chains, black nodes represent observed cases of TB; gray represents unobserved. Arrows represent transmission events. A, Perfect surveillance when all cases originating from a single index case are observed without censoring. B, Incomplete ascertainment when i missing cases result in a cluster size of $Y - i$. C, Broken chains from incomplete ascertainment, such that the position of the missing case in the chain results in pseudo-clusters. D, Censored chains are ongoing chains at the time of data collection. E, Overlapping chains result when n chains are unable to be disentangled, as described in the methods, resulting in a single cluster of size y with n chains.



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.0	2.72	0.96	0.62	0.45	0.36	0.29	0.24	0.20	0.17	0.15
0.9	2.11	0.89	0.59	0.43	0.34	0.28	0.23	0.20	0.17	0.15
0.8	1.89	0.83	0.54	0.41	0.32	0.27	0.23	0.20	0.17	0.15
0.7	1.56	0.76	0.52	0.38	0.31	0.26	0.22	0.19	0.17	0.15
0.6	1.45	0.71	0.47	0.36	0.29	0.25	0.21	0.19	0.17	0.15
0.5	1.17	0.64	0.45	0.33	0.28	0.24	0.21	0.18	0.16	0.15
0.4	1.04	0.56	0.40	0.31	0.26	0.22	0.20	0.18	0.16	0.15
0.3	0.87	0.50	0.36	0.29	0.24	0.21	0.19	0.17	0.16	0.15
0.2	0.72	0.42	0.32	0.26	0.23	0.20	0.18	0.17	0.16	0.15
0.1	0.55	0.35	0.28	0.23	0.21	0.19	0.18	0.17	0.16	0.15

FIGURE 2.

Estimates of k by case ascertainment probabilities. Bias from missing TB cases through simulated passive and active surveillance modeled under the empirical estimates of R = 0.50 and k = 0.15. Passive surveillance (p_1) represents the proportion of TB cases that were properly ascertained (e.g., correctly diagnosed, cultured, and genotyped). Active surveillance (p_2) represents the proportion of otherwise undiagnosed cases that were ascertained because of additional public health efforts (e.g., contact tracing) triggered after at least one other case in the chain was observed through passive surveillance (see methods). Numbers in the center of each combination of p_1 and p_2 represent the median value of k from 500 simulated surveillance systems for each combination of p_1 and p_2, as described in the methods. Additional R and k values shown in eFigure 2A, B (http://links.lww.com/EDE/B887).

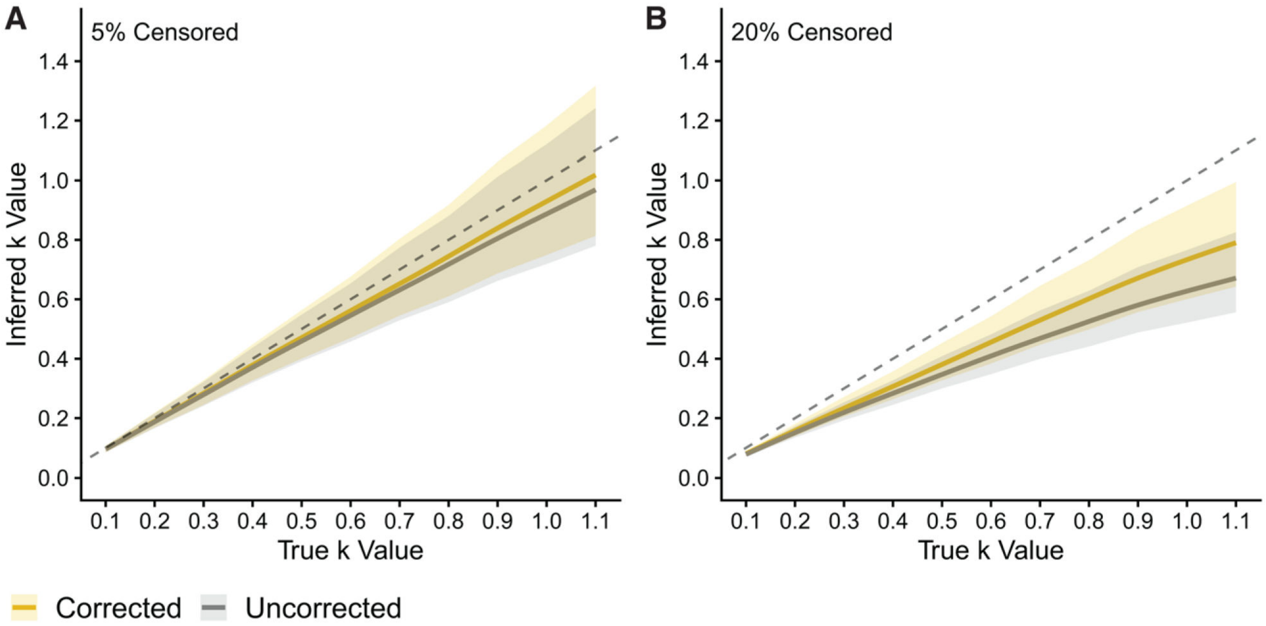


FIGURE 3.

Impact of censoring on estimates of k. Results of simulated TB surveillance systems censored according to the methods, for each value of K between 0.1 and 1.1 in 0.01 increments ($R = 0.90$). A, 5% of clusters are censored ($p_{cens} = 0.05$). B, 20% censoring ($p_{cens} = 0.20$). “Corrected” indicates estimates when accounting for censoring in the likelihood (i.e., $P(Y \geq y)$). “Uncorrected” indicates censoring was not accounted for in the likelihood equation (all $P(Y = y)$). The dashed line represents perfect inference (e.g., $\hat{k} = k$). Solid lines represent the median of 500 simulated surveillance systems for each K value; shading represents 95% confidence intervals. Additional censoring values shown in eFigure 3 (http://links.lww.com/EDE/B887).

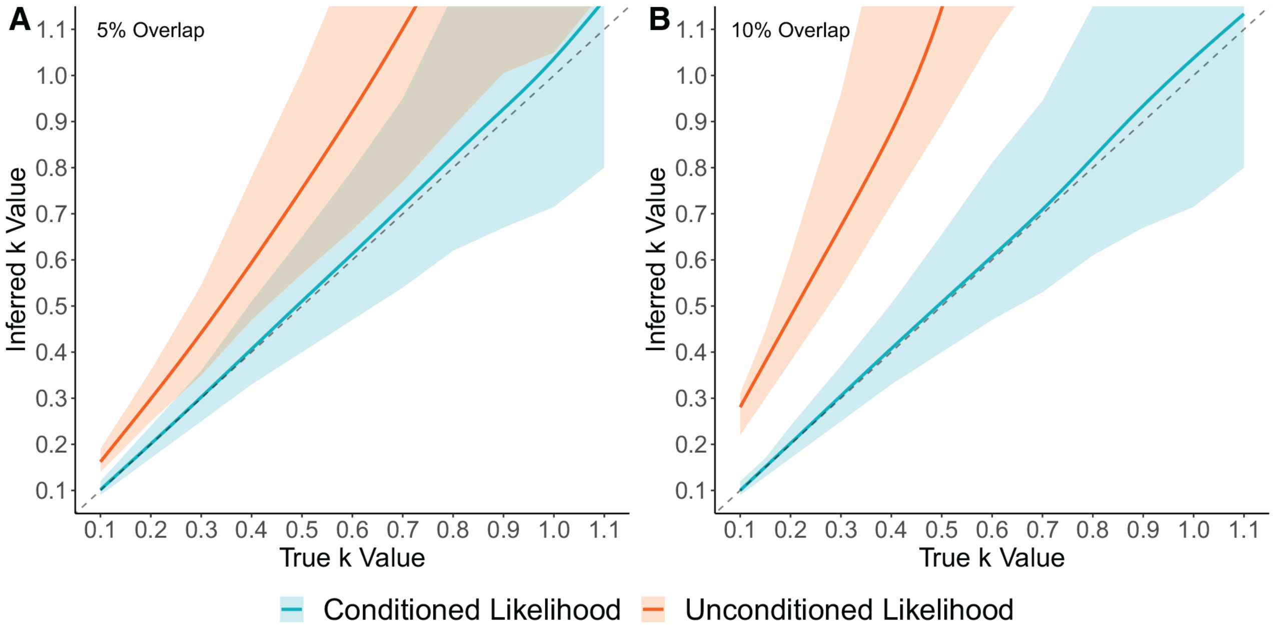


FIGURE 4.

Bias in $\hat{k}$ arising due to overlapping clusters. Results of simulated TB surveillance systems with overlapping chains according to the methods, for each value of k between 0.1 and 1.1 in 0.01 increments ($R = 0.50$). Simulations with (A) 5% and (B) 20% of TB transmission chains in the surveillance system overlap, per methods ($p_{cens} = 0.05$ and 0.20, respectively). “Conditioned” likelihood shows results when the probability is conditioned on the number of overlapping chains in a given cluster (i.e., $P(Y = y|n)$); “Unconditioned” shows results when the number of overlapping chains is ignored (i.e., $P(Y = y)$). Shaded areas indicate 95% confidence intervals; dashed line represents perfect inference $\left(\hat{k} = k\right)$. Additional p_{cens} values shown in eFigure 4 (http://links.lww.com/EDE/B887).

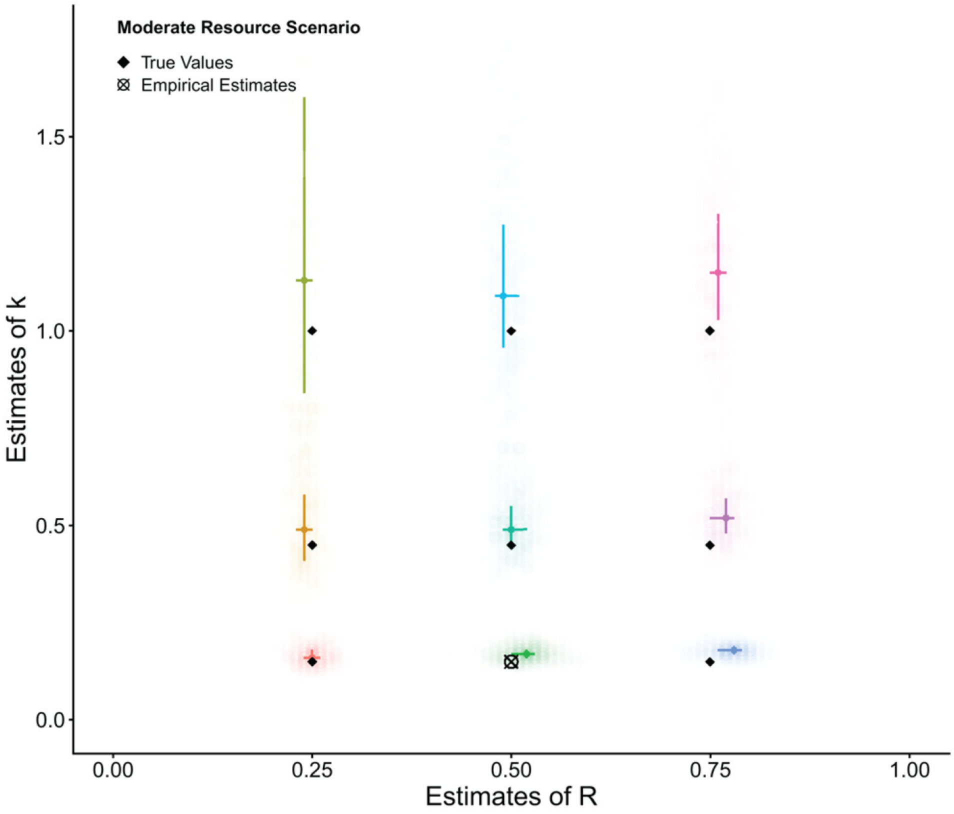


FIGURE 5.

Performance of inference procedure under imperfect surveillance. Results of 500 simulated TB surveillance systems under combined imperfect surveillance scenarios under moderate-resource assumptions. Lines represent the interquartile range for each R and k combination; dots represent median values. Diamonds represent true values. R values were simulated at 0.25, 0.50 (empirical), and 0.75. k values were simulated at 0.15 (empirical), 0.45, and 1.0. Analogous figures for high- and low-resource scenarios are available in eFigure 5 (http://links.lww.com/EDE/B887).

TABLE 1.

Comparison of Individual- and Cluster-level Methods of Parameter Inference

		True k = 0.25		True k = 0.50		True k = 0.75	
True R	**MLE**	**Individual Data**	**Cluster Data**	**Individual Data**	**Cluster Data**	**Individual Data**	**Cluster Data**
0.90	$\widehat{R}$	0.90 (0.89, 0.91)	0.90 (0.89, 0.91)	0.90 (0.89, 0.91)	0.90 (0.89, 0.91)	0.90 (0.89, 0.91)	0.90 (0.89, 0.91)
	$\widehat{k}$	0.25 (0.24, 0.26)	0.25 (0.24, 0.26)	0.50 (0.49, 0.51)	0.50 (0.47, 0.53)	0.75 (0.74, 0.76)	0.75 (0.70, 0.81)
0.70	$\widehat{R}$	0.70 (0.69, 0.71)	0.70 (0.69, 0.71)	0.70 (0.69, 0.71)	0.70 (0.69, 0.71)	0.70 (0.69, 0.71)	0.70 (0.69, 0.71)
	$\widehat{k}$	0.25 (0.24, 0.26)	0.25 (0.24, 0.27)	0.50 (0.49, 0.51)	0.50 (0.47, 0.54)	0.75 (0.73, 0.78)	0.76 (0.70, 0.82)
0.50	$\widehat{R}$	0.50 (0.49, 0.51)	0.50 (0.49, 0.51)	0.50 (0.49, 0.51)	0.50 (0.49, 0.51)	0.50 (0.49, 0.51)	0.50 (0.49, 0.51)
	$\widehat{k}$	0.25 (0.24, 0.26)	0.25 (0.23, 0.27)	0.50 (0.48, 0.52)	0.50 (0.47, 0.55)	0.75 (0.72, 0.80)	0.74 (0.68, 0.83)
0.30	$\widehat{R}$	0.30 (0.29, 0.31)	0.30 (0.29, 0.31)	0.30 (0.29, 0.31)	0.30 (0.29, 0.31)	0.30 (0.29, 0.31)	0.30 (0.29, 0.31)
	$\widehat{k}$	0.25 (0.24, 0.27)	0.25 (0.23, 0.27)	0.50 (0.46, 0.54)	0.50 (0.44, 0.56)	0.75 (0.69, 0.82)	0.74 (0.65, 0.86)

We compared parameter inference using the cluster-based procedure (“cluster data”) with classical methods that utilize the full distribution of individual secondary cases (“individual data”). Values represent the median maximum likelihood estimates (MLE [interquartile range]) from 500 simulated surveillance systems, each originating with 2,000 individual chains, under perfect TB surveillance (see Methods).

TABLE 2.

Simulated Combined Imperfect Scenarios

	Parameter Values		
Surveillance	**High Resource**[37,38,40]	**Moderate Resource**[36,37,41]	**Low Resource**[36,42]
Proportion of cases identified via. passive surveillance (P_1)	0.90	0.75	0.50
Additional cases identified via. active surveillance (p_2)[a]	0.75	0.50	0.25
Proportion of clusters censored (p_{cens})	0.05	0.10	0.10
Proportion of clusters overlapping (e.g., with 2 or more index cases) (p_{clust})	0.15	0.20	0.20

Model parameters were assigned to simulate the extent of imperfect surveillance in three theoretical scenarios representing high-resource, moderate-resource, and low-resource TB surveillance systems.

[a] Not applicable to wholly unobserved transmission chains; only unobserved cases in transmission chains where at least one case in the chain was observed to trigger active case finding procedures were subject to p_2.

TABLE 3.

Estimates of Transmission Parameters R and k for TB Transmission in the United States by Timeframe and Geographic Definition of Clusters

		Primary Definition		Alternative Definitions ($\hat{k}$)			
		WGS and Epidemiological Link Data Used to Identify n Values, All $Y \geq 10$ Assumed Censored		All $n = 1$	All $n = y$	No Censoring	All $Y > 3$ Censored
Sampling Timeframe	**Geographic Catchment**	$\hat{R}$ **(95% CI)**	$\hat{k}$ **(95% CI)**	$\hat{k}$ **(95% CI)**	$\hat{k}$ **(95% CI)**	$\hat{k}$ **(95% CI)**	$\hat{k}$ **(95% CI)**
5 years (2012–2016)	County	0.17 (0.16, 0.17)	0.09 (0.08, 0.09)	0.09 (0.09, 0.10)	0.12 (0.11, 0.13)	0.09 (0.09, 0.10)	0.07 (0.07, 0.08)
	State	0.27 (0.26, 0.28)	0.11 (0.11, 0.12)	0.12 (0.11, 0.13)	0.16 (0.15, 0.17)	0.12 (0.11, 0.13)	0.08 (0.07, 0.09)
3 years (2014–2016)	County	0.14 (0.13, 0.15)	0.08 (0.07, 0.09)	0.08 (0.07, 0.09)	0.10 (0.09, 0.11)	0.08 (0.07, 0.09)	0.06 (0.06, 0.07)
	State	0.23 (0.22, 0.24)	0.11 (0.10, 0.11)	0.11 (0.10, 0.11)	0.14 (0.13, 0.15)	0.11 (0.10, 0.12)	0.08 (0.08, 0.09)

Our primary approach defined the number of transmission subclusters, n, within large genotypic clusters, Y, based on WGS and epi-link data. Sensitivity analyses varied definition assumptions: $n = 1$ or $n = y$ for all large clusters ($Y \geq 10$), no censoring, and all clusters larger than 3 were censored.