Proceedings of the 2014 Winter Simulation Conference
A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, eds.

ESTIMATING THE PROPORTION OF TUBERCULOSIS RECENT TRANSMISSION VIA SIMULATION

Parastu Kasaie
David W. Dowdy

Department of Epidemiology
Bloomberg School of Public Health
The Johns Hopkins University
Baltimore, MD 21205, USA

W. David Kelton

Department of Operations, Business Analytics, and Information Systems
University of Cincinnati
Cincinnati, OH 45221-0130, USA

ABSTRACT

Tuberculosis (TB) is an infectious disease that can progress rapidly after infection or enter a period of latency that can last many years before reactivation. Accurate estimation of the proportion of TB disease representing recent versus remote (long ago) transmission is critical to disease-control policymaking (e.g., high rates of recent transmission demand more aggressive diagnostics). Existing approaches to this problem through cluster analysis of TB strains in population-based studies of TB molecular epidemiology are crude and prone to bias. We propose an agent-based simulation of TB transmission in conjunction with molecular epidemiologic techniques that enables study of clustering dynamics in relation to disease incidence, diversity of circulating strains, sampling coverage, and study duration. We perform a sequence of simulation experiments with regard to different levels of each factor, and study the accuracy of estimates from the cluster-analysis method relative to the true proportion of incidence due to recent transmission.

1 INTRODUCTION

Tuberculosis (TB) infection is associated with a long (unknown) latency period during which individuals are neither symptomatic nor infectious, and are subject to a time-varying risk of progression to active disease. Unlike some other airborne diseases, however, available diagnostic techniques do not allow for differentiating the original timing of TB transmission in diagnosed cases, and the true proportion of incidence due to recent transmission (in contrast to reactivation of remote infections that originated many years ago) remains uncertain. Accurate estimation of this value is especially important for disease-control policy-making and choice of strategies focusing on recent versus remote infections – in areas where reactivation from remote infection is important, preventive therapy may be favored, whereas in areas where recent infection accounts for most of the disease, aggressive diagnostics and therapeutic interventions may have more impact. In this setting, *Cluster analysis* of TB strain types offers an indirect approach for estimating the proportion of disease attributable to recent transmission as opposed to progression from long-ago infection that has been latent in the individual.

Molecular studies of TB use a variety of molecular subtyping techniques to characterize specific strains of TB on the basis of DNA patterns. Such techniques take advantage of the existence of DNA elements in the genome of *M. tuberculosis* (MTB) that differ in number and genomic location from one strain to the next, but tend to mutate at a sufficiently slow rate that a "fingerprint" can be identified that is identical between people who are connected by a small number of transmission events. DNA fingerprinting techniques – which consist of isolating mycobacterial DNA and evaluating the number and location of these repetitive elements in the laboratory – are widely used in documenting the transmission of MTB strains across populations, and with the improvement of DNA sequencing techniques, their ability to discriminate recent versus remote transmission is becoming much stronger. Comparison of patients' isolates during

previous TB outbreaks suggests that patients with similar DNA patterns were infected from a common source and therefore constitute an epidemiologically linked cluster (e.g., sharing a similar (/unrelated) strain type between two individuals provides evidence for (/against) transmission of disease from one to another) (Chevrel-Dellagi et al. 1993). Moreover, given the short period of disease development among these patients (due to higher risk of disease development in the first five years of the latency period), the *clustering level* (i.e., the proportion of TB cases that share identical DNA patterns) can be used as a general identifier of the extent of disease attributable to recent TB infection.

To illustrate the computational steps, consider a sample of size N from a population, organized into n clusters. We represent each cluster by its size (c), with $1 \leq c \leq C$. A cluster of size one (c =1) represents an isolate of a unique DNA pattern that is not clustered with any other sample isolates. In general, a cluster of size c includes c individuals with identical DNA fingerprints. Furthermore, we let n_c denote the frequency of each cluster in the population. Based on previous findings, the resulting clusters can be interpreted as epidemiologically linked chains of recently transmitted disease, and unique isolates are cases of reactivation (Murray 2002b). The proportion of clustered cases in the sample can be used as an estimate of the ratio of incidence due to recent transmission. This value can be further adjusted by assuming one case as the source of infection in each cluster (which may as well be due to reactivation) and removing it from the calculations. The procedure is referred to as the "n-1" cluster-analysis method in which the total number of secondary cases (i.e., clustered cases minus the source cases) is used to estimate the clustering-level attributable to the rate of recent transmission as $\sum_{c=1}^{C}(c-1)\times n_c / N$.

Despite wide application of the *"n-1"* cluster-analysis approach to distinguish between remote and recent TB infection, questions remain regarding the validity of this approach in various settings and the accuracy of estimates for the true level of recent transmission. Previous studies have identified a number of factors affecting the performance of this approach for estimating the true level of clustering in the population, and investigated the effects in a number of data-driven settings (Vynnycky et al. 2001; Murray 2002b). Nevertheless, the relationship of such factors in the presence of each other and their overall effect in a general domain remains uncertain. Specifically, estimates can be biased by the coverage of molecular fingerprint data (i.e., number of TB cases with available isolates for fingerprinting), duration of time over which molecular fingerprints were obtained, underlying TB incidence in the population, and degree of genetic diversity of *M. tuberculosis* in the population.

Since the underlying cluster-size distribution is uncertain and therefore less amenable to an analytical approach, we develop an agent-based simulation of TB transmission that models the clustering dynamics, and generates synthetic corresponding DNA repositories of TB patients over time. Subsequently, we perform a sequence of simulation experiments with regard to levels of each factor and estimate the rate of recent transmission using the cluster-analysis method. This allows us to study the impact of each factor on the final estimates of the clustering level and the accuracy of the cluster-analysis estimate relative to the true proportion of incidence due to recent transmission.

2 BACKGROUND

Previous studies have suggested that underlying distribution of TB cluster sizes affects the final estimate of the proportion of cases due to recent transmission. Variation in the distribution of TB clustered isolates in different populations is assumed to reflect different TB transmission dynamics and intensities, such that a high proportion of large clusters suggests ongoing new TB transmission, whereas a dominance of unique cases implies wide reactivation of infections (likely originated long ago) without further spread. From a computational perspective, this distribution represents the collection of cluster-size frequencies as a function of time (i.e., $n_c(t)$, c = 1,…, $C(t)$), and can be considered as a complex (unknown) function of disease-transmission dynamics, social mixing patterns, pathogen-related characteristics, etc. (Murray 2002a). Moreover, study design and sampling strategy for acquiring representative samples of a population's molecular fingerprint data are key factors affecting the precision of clustering-level estimates. Sampling coverage and duration determine the likelihood of capturing clustered cases involved in the same

chain of transmission, and subsequently affect the final estimate of the clustering level (e.g., sampling studies with low coverage or short duration tend to miss clustered cases in the same chain of transmission, and therefore underestimate the level of clustering). Glynn et al. (1999) use stochastic simulation models based on real and hypothetical populations and demonstrate the influence of incomplete sampling on the estimates of clustering obtained. Murray (2002b) proposes an analytical model for estimating the magnitude of bias introduced by incomplete sampling. However, while Murray's models describe the bias inherent in the classical "$n - 1$" cluster analysis, they do not provide an easy mechanism for translating clustering data into better estimates of recent transmission. Another study models the role of age structure and study duration on the relationship between clustering and the ratio of incidence due to recent transmission in a data-driven setting for The Netherlands (Vynnycky et al. 2001), and suggests the impact of these factors on interpretation of clustering levels to differentiate recent versus remote infection patterns.

Such findings imply that estimates of recent transmission obtained by molecular-analysis methods cannot be compared across studies using different sampling coverage or in which the distribution of cluster sizes is expected to vary. We address this problem by conducting wide sensitivity analyses of the potential estimation error with regard to underlying TB incidence, degree of genetic diversity of *M. tuberculosis*, and sampling characteristics. Since the true distribution of cluster sizes is a matter of uncertainty, we propose an agent-based simulation of TB to model the dynamics of disease diffusion and clustering over time. The model is used to generate simulated DNA repositories of TB patients in a variety of scenarios. The repositories further allow for explicit implementation of sampling studies with various levels of coverage and duration.

3 METHOD

We propose a simulation-based approach to study the performance of the cluster-analysis method in estimating the rate of TB recent transmission, and explore the role of various factors in relation to the epidemic system and sampling strategy. We develop an agent-based simulation of TB transmission to model the long-term dynamics of disease diffusion and strain clustering. The simulation model lies at the heart of our experimental framework, and is used to generate a virtual repository of population DNA fingerprints (corresponding to TB strain types) under different scenarios. This repository represents the dynamic distribution of strain types over time, and is further investigated through a series of sampling experiments with regard to combinations of study coverage and duration. In this section, we present the general characteristics of the TB simulation model, discuss our strategy for modeling disease strain diversity, introduce the experimental factors, and illustrate the simulation-experiment framework for the analysis.

3.1 An Agent-Based Simulation Model of a TB Epidemic

We develop an agent-based simulation (ABS) model, coded in C++, of a TB epidemic in an age-structured population with homogenous mixing. This model takes a population of individuals described by their TB disease state and specific MTB strain. Once a person enters a health state (e.g., LLTB), possible exit routes are identified (e.g. slow progression to active TB (ATB)) and the duration of time before existing through that route is generated from a corresponding geometric distribution. At the end of each timestep, the code checks the current model time against the remaining time to exit each state, and executes events that are due. We choose a timestep of one month for our analysis, which is long enough to facilitate crude estimation of annual strain clustering, and short enough to reveal the dynamics of TB transmission across a year-long period. Disease progression and individual contacts are modeled at the end of each month, while the birth/death process and migration events are modeled at the end of each year.

Since the goal of modeling is simulating the dynamics of disease transmission and strain-clustering in a general setting and over a long period of time, we do not rely on historical data from specific TB outbreaks, and instead adopt regularly cited parameter values from previous modeling studies (Section 3.1.2). Furthermore, we validate the simulation by aligning the aggregate outputs (e.g., incidence and mortality) against a corresponding deterministic model of transmission in a number of sample scenarios (using a procedure similar to that discussed in (Kasaie et al. 2013)).

3.1.1 Population Structure and Contact Network

We consider an age-structured homogeneous population of 100,000 individuals with an initial uniform age distribution (0-90 years old). Individuals are subject to an age-specific annual mortality rate, and die at a maximum age of 90 years. Deceased individuals are replaced with a number of newborns at the end of each year such that the average population size remains constant over time. The annual number of newborns is therefore generated from a non-stationary Poisson process with a mean value set to the total number of people dying during each year.

The initial model assumes a homogeneous mixing structure in which all members of the population have an equal chance of contact with each other. The contact events are modeled at the end of each month and for computational efficiency, only the infectious contact events are modeled (i.e., contacts among infectious cases and the rest of the population). We assume that the individual's number of contacts per timestep (one month) is generated from a Poisson distribution with mean of λ. The value of λ (the contact rate) is subsequently tuned to calibrate the average incidence rate at steady state.

The simulation model starts from an existing endemic calibrated to the equilibrium level of a corresponding deterministic transmission model. We assume that all initially infected individuals (in non-susceptible states) carry unique strains of TB. A disease strain can be transferred from an infectious person to another individual upon a successful infectious contact leading to disease transmission. Infected individuals carry a single and only the latest acquired TB strain at each point of time (i.e., in case of a reinfection, the new strain replaces the previous strain). Since the true distribution of TB strain types is not known, the simulation model is used to trace the dynamics of transmission, strain transfer, and subsequent cluster formation through time. In the current model, we assume no error in the fingerprinting method; that is, we assume that all cases of TB linked by transmission events will be correctly identified as belonging to a cluster of individuals with the same TB strain. For example, cluster analysis of the strains in Figure 1 would compute the total number of secondary cases of ATB as 10 (1+2+3+4), number of active strains as 20 (N), and the proportion of incidence due to recent transmission as 50%).

3.1.2 TB Natural History

TB natural history is modeled at an individual level using five main TB health states as shown in Figure 1. A person is assumed to be born in full health and susceptible to TB. Upon successful transmission of the disease, the person enters the early latent TB (ELTB) state for a period of five years. During this time, individuals are at high risk of progression to ATB (primary infection); this probability declines with each year in the ELTB state (i.e., the ELTB state actually consists of five sub-states, each lasting one year), according to estimates of adult TB progression following infection (Vynnycky & Fine 1997). After five years post-infection, individuals enter the late latent TB (LLTB) state, when they have a very low but non-zero probability of progression to active disease in each year (reactivation). Reinfection can occur during early or late latency, and is modeled as a return to year one of the ELTB state; however, individuals with latent TB are assumed to have a degree of immunologic protection from reinfection, such that the probability of infection for an active-to-susceptible contact is higher than the probability of reinfection for an active-to-latent contact.

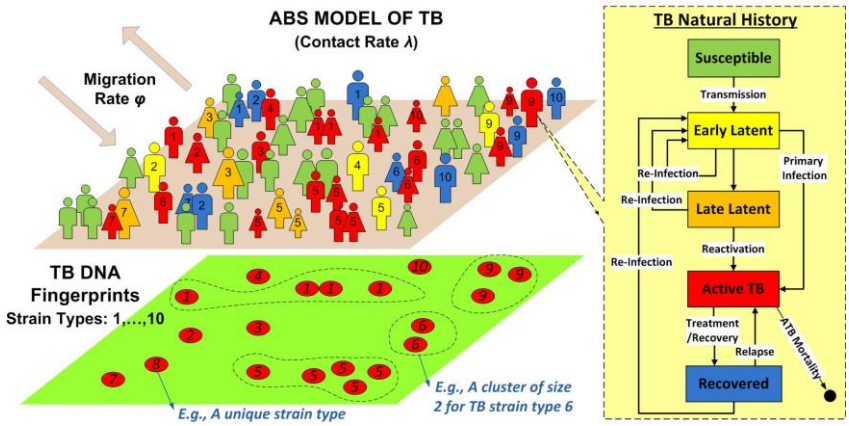

Figure 1: ABS Model Outline

Patients with ATB are infectious and subject to an increased probability of mortality. Infectiousness is modeled as zero at the start of the active period, increasing thereafter as a linear function of time for the first nine months of the disease (i.e., as the bacillary burden grows), and stable thereafter at the maximum level until the individual is diagnosed and treated, or dies. This duration of disease has previously been shown to replicate the prevalence/incidence ratio as estimated by the World Health Organization (Dowdy et al. 2013). The maximum transmission probability is assumed to be one (full infectiousness), thereby reflecting a combination of intrinsic transmissibility and external factors (e.g., crowding, poverty) that may modify the probability of effective TB transmission. Upon diagnosis and initiation of effective treatment, or alternatively through spontaneous resolution, individuals with active TB enter the non-infectious recovered state. We calibrate the treatment/recovery rate such that individuals remain infectious for an average of 11 months prior to treatment, as estimated by the WHO (World Health Organization 2013b). Moreover, recovered individuals are subject to a risk of disease relapse in the first two years of recovery. The annual rate of relapse (in the first two years) is tuned to calibrate the proportion of incidence among previously treated individuals. These individuals are also subject to a risk of reinfection, as described above.

Population characteristics and disease parameters are calibrated to the literature for settings of medium-to-high TB incidence, as provided in Table 1. We consider a baseline scenario for a setting of medium TB incidence (200 per 100,000/year, representing India) and determine the simulation transient period to be 100 years (using a method similar to that in Kasaie et al. (2013)). The value of λ (contact rate) is subsequently tuned to calibrate the average incidence rate at steady state (starting in year 101).

Table 1: Simulation Model Parameters and Values

Parameter	**Value (Description/Source)**
Natural mortality rates	Age-specific mortality rates for India (World Health Organization 2013a)
ATB mortality rate	0.12 per year (Tiemersma et al. 2011)
Annual primary infection rate (year 1 to 5 post infection)	(8.66, 3.55, 1.12, 0.74, 0.24)% (Vynnycky & Fine 1997)
Reactivation rate	0.001 (Dowdy & Chaisson 2009)
Treatment/Recovery	0.9 per year (Providing disease duration of 11 months)
Relapse rate (first two years)	0.06 per year (Providing 15% incidence among previously treated individuals)
Latent immunity toward re-infection	50% (Andrews et al. 2012)

3.1.3 Simulating TB Strains

A critical challenge in developing a simulation model of TB strain clustering is the maintenance of appropriate genetic diversity. Due to clustering effects, the initial strain diversity will not persist and the number of circulating strains in the population falls quickly as the model proceeds. Although mutation rates of MTB are remarkably low, TB strains circulating in real human populations (e.g., those of the United States (Moonan et al. 2012)) are quite diverse, presumably owing to the high historical incidence of TB more recently sustained by population growth despite overall declines in per-capita incidence. These historical dynamics are difficult to replicate because they occur on the order of centuries, and no reliable data on such trends exist. Thus, to construct a simulation model of TB with appropriate strain diversity, alternative methods to a simple "mutation rate" over time must be employed.

One alternative approach for modeling long-term diversity of disease strains (in a single population) assumes mixing of individuals with an external population of unlimited genetic diversity. A convenient approach in modeling such mixing involves a migration process between the study population and the external population (with equal rates φ to maintain a fixed population size), such that each immigrant who is TB-infected carries a new (and unique) strain of TB. Figure 2 shows the number of surviving TB strains over time under different levels of the migration rate (φ) in a population of 100,000 people with incidence of 200/year. With no migration (closed population with internal births and deaths), the number of strains maintained within the population falls to a small number within approximately two lifetimes (Figure 2, darkest blue line at the bottom). In contrast, once the migration rate reaches approximately 10% of the population per year, the number of circulating strains reaches an equilibrium level that approximates the number of people in the population who are latently infected with TB.

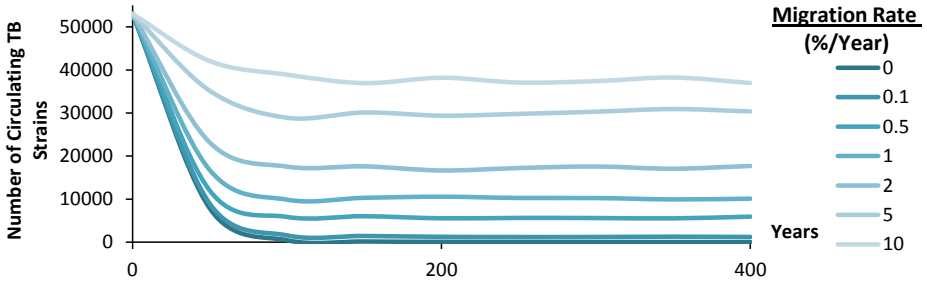

Figure 2: Pattern of TB Strain Survival in Relation to Migration Rate

The migration rate (φ) is subsequently tuned to calibrate the number of surviving strains through time. In order to assist with the computational procedure, we match new immigrants to the population with random members of the population and update their TB strain type upon re-entering the population (e.g., randomly selected individuals are chosen to leave the population, and each one is replaced with another randomly selected member of the population carrying a unique TB strain if infected). The implicit correlation among individuals' states as induced by this procedure is in turn beneficial to our analysis by reducing the variability of disease equilibrium-state levels, and population age structure. The migration process takes place at the end of each year, and the actively infected immigrants entering the population are not counted toward disease incidence (since they acquired the disease prior to entering the population).

3.2 Experimental Factors

We consider four factors in relation to the underlying disease incidence, strain diversity, and sampling characteristics. Given the uncertainty regarding the true cluster-size distribution, we rely on our simulation-model predictions for the dynamics of TB transmission and survival pattern of TB strains over time. To describe the expected level of clustering in relation to transmission dynamics, we consider the TB incidence level as the first experimental factor and explore a range of medium-to-high TB incidence levels in our analysis. The incidence level describes the annual number of people developing ATB each year and

is calibrated by tuning the contact-rate parameter λ in the simulation model. Moreover, we define the *Ratio of Circulating TB Strains* (RCS) as a measure to describe the diversity of TB strain types over time. Since the crude number of circulating TB strains is a function of population size and underlying incidence, we standardize this value with regard to the size of the TB-infected population at each point in time, and consider RCS as the second experimental factor:

$$RCS(t) = \frac{\text{Number of circulating strains at time } t}{\text{Non-susceptible population size at time } t}$$

RCS levels range between 0 and 1, where a value of 1 corresponds to a population with no DNA clusters such that every infected person (whether latent, active, or recovered) carries a unique strain of TB. Accordingly, the average RCS level is calibrated by tuning the migration rate φ. Due to the unknown level of RCS in practice, we explore a wide range of variation corresponding to migration rates ranging from 0.1% per year (representing relatively closed populations) to 10% per year (representing dynamic population exchange).

Moreover, we assume that TB patients are recruited in the study upon receiving treatment and recovery, and describe the sampling strategy with regard to sampling duration (d) denoting the length of the sampling period, and sampling coverage (p) denoting the probability of a successful case ascertainment. We consider eight levels of variation for study duration, equally distributed over a range of 5 to 40 years of sampling, and five levels of variation for the sampling coverage ranging from 0 to 1. Table 2 summarizes the information on selected experimental factors and their levels. In summary, we consider 8 (incidence levels) × 8 (migration levels) × 8 (sampling duration levels) × 5 (sampling coverage levels) = 2,560 unique parameter sets in our experiments. In these simulations, a sampling study with a duration of 5 years and coverage of 20% represents the highest incompleteness and thus misses the largest portion of TB patients (i.e., a large number of TB patients are not sampled and are excluded from the cluster analysis).

Table 2: Experimental Factors and the Levels of Each Factor Considered in Different Experiments

Factors	**Levels**
Incidence (/year)	(100, 150, 200, 250, 300, 350, 400, 450)
Migration Rate (%)	(0.1, 0.5, 1, 2, 4, 6, 8, 10)
Sampling duration (years)	(5, 10 15 20 25 30 35 40
Sampling coverage	(0.2, 0.4, 0.6, 0.8, 1)

3.3 Simulation Experiment Framework

We design the simulation experiment framework in two levels (Figure 3). In the first level, we consider various scenarios corresponding to combinations of incidence and RCS levels. Each simulation scenario, $S_{i,}$ is associated with a unique set of input parameters (contact rate i and migration rate j, $i,j = 1,\ldots,8$) and requires a fresh startup. We record the DNA fingerprints (e.g., strain types) of all treated TB patients at the end of each year, and save them into a corresponding DNA repository representing the dynamic distribution TB strain types over time. The repositories are subsequently used in the second level of analysis for modeling various sampling experiments corresponding to combinations of study coverage and duration levels ($e_{p,d}$, $p = 1,\ldots,5$, $d = 1,\ldots,8$). Each experiment is carried out over a random selection from the DNA repository (for d consecutive years), recruiting a random sample of populations (with probability of ascertainment p) , and estimating the clustering level using the *"n-1"* method (attributable to the rate of recent transmission). All scenarios are run for R replications, and the sampling experiments are carried over all the generated DNA repositories ($8\times8\times8\times5\times R$ observations). Figure 3 shows the conceptual model of simulation experiments for this analysis.

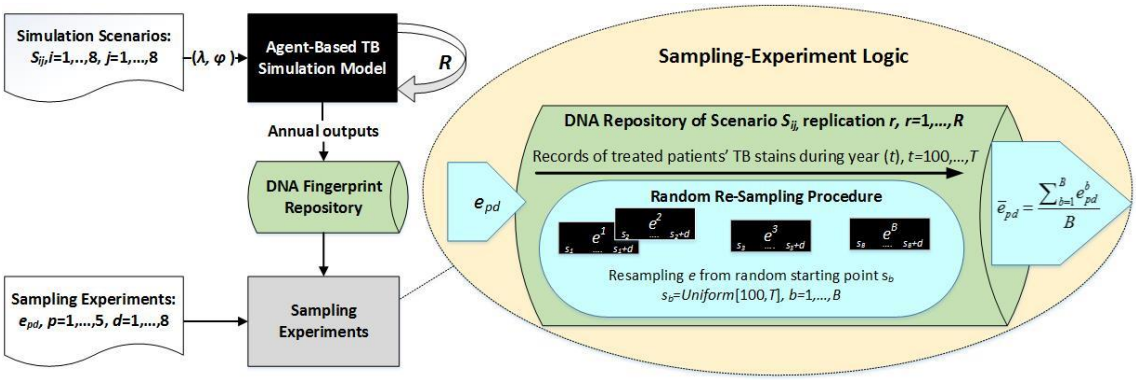

Figure 3: Conceptual Model of Simulation Experiments.

Sampling-Experiment Logic: We use a repeated-sampling strategy to increase the precision of the sampling experiments' results carried over a DNA repository. Consider a single replication r of a scenario $S_{i,j}$, where r =1,…,R and i,j = 1,..,8. The TB isolate repository corresponding to this replication includes the annual TB isolates of all treated patients in this model: from the beginning of the simulation model to year T (end of the simulation). Next, consider a sampling experiment $e_{p,d}$, where p = 1,…,5, d = 1,…,8. This experiment is associated with a pre-specified level of duration and coverage (p,d) and will be carried over the DNA fingerprint repository as follows: beginning from year s, we sample the strain types of recovered patients with a likelihood of p, and continue the sampling procedure for d years. Collected strain types are then analyzed using an *"n-1"* cluster-analysis method as described in Section 2, and the estimated level of clustering from this method is compared to the actual simulated level of recent TB transmission.

A result collected in this fashion, however, is subject to stochastic variation of TB dynamics throughout the simulation period s to s+d, which eventually inflates the variability of the experiment's results and reduces the power of design for differentiating the underlying variance from meaningful factor effects. Therefore, we use a random re-sampling procedure to increase precision (Figure 3). We repeat each experiment $e_{p,d}$ for B times, each time using a random starting point s, and report the average results. The sampling start time s is chosen randomly and uniformly from the beginning of simulation steady state (year 101) to year T-d. The choice of B is made such that the relative precision of the final reported results is at worst 5% . In this analysis, we consider R = 10 replications and assume a time horizon of T = 1000 years, which provides 900 years of annual TB DNA fingerprint data for our sampling experiments (i.e., skipping the initial 100-year transient period in each model).

4 RESULTS

Analysis of Simulation Scenarios: The simulation scenarios are characterized by variation in underlying incidence and RCS levels. The average incidence level is calibrated by tuning the value of the contact rate parameter λ in the simulation model, and is a rather close linear function of this parameter's value across all scenarios, regardless of the migration rate (Figure 4, left panel). Thus, the migration rate φ enables changing the diversity of circulating strains (Section 3.3.2) and tuning the average RCS level at equilibrium without affecting incidence. The RCS level, however, further depends on the underlying incidence level: a higher incidence level corresponds to faster transmission dynamics, which in turn increases the proportion of TB strains that are clustered, thereby mildly lowering the average RCS level in the long term (Figure 4, right panel).

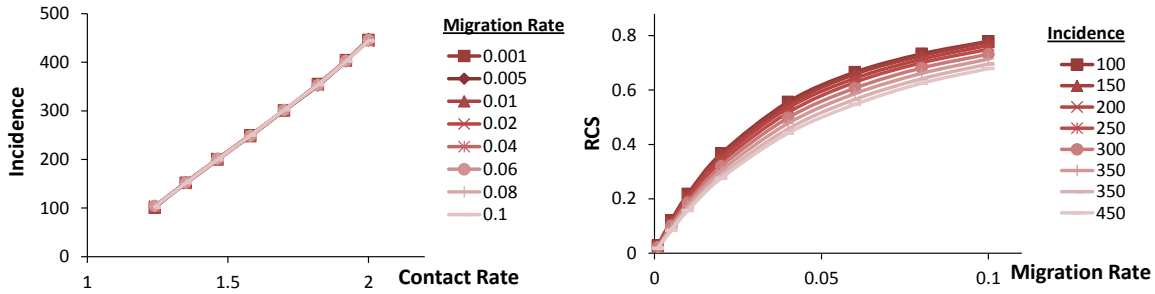

Figure 4: Calibrating the Simulation Scenarios.

Moreover, our analysis for monitoring the timing of infections across incident cases suggests that higher incidence levels are associated with a higher proportion of disease due to recent transmission (primary infection). Figure 5 shows the proportion of incidence due to primary progression, reactivation from "late latent" infection, and relapse from prior recovered TB at varying levels of incidence across simulation scenarios, while all the other model parameters are fixed. The pattern explains the role of ELTB state in higher transmission settings, which in turn increases the cumulative risk of fast progression.

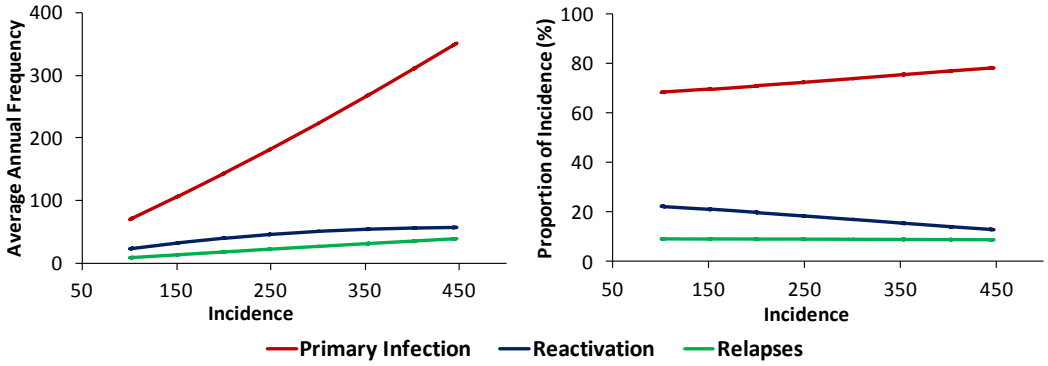

Figure 5: Pattern of Changes Across Sources of Infection.

Analysis of Sampling Experiments: We next consider the sampling experiments corresponding to various levels of sampling duration and coverage. Each experiment is carried over all DNA repositories associated with replications of simulation scenarios. Since the goal of the analysis is studying the precision of the *"n-1"* cluster-analysis method in estimating the proportion of incidence due to recent transmission, we consider estimation bias as the main outcome of interest and compute this quantity in each scenario as: (proportion of TB estimated by *"n-1"* method to represent recent transmission) – (true proportion of incidence due to recent transmission). Thus, a high positive estimation bias corresponds to an estimate that is much higher than the simulated "truth," whereas a very negative estimation bias corresponds to an estimate that is much lower. Figure 6 depicts the estimation bias across a number of simulation scenarios defined by underlying incidence, ratio of circulating strain number to total TB cases (RCS), duration of sampling in years (*d*), and population coverage of the sampling technique (*p*). The graphs are presented in increasing order of incidence (from left to right), and increasing order of average RCS (top to bottom), with light blue corresponding to an unbiased estimate from the *"n-1"* method.

Increasing the RCS level at a fixed incidence rate (i.e., moving down each column) reduces the overestimation bias of the *"n-1"* method for sampling schemes with good coverage (high *p* and *d* values) but worsens *"n-1"* estimates from smaller-scale studies (which at low RCS values reflect compensating biases of missed clustering in one direction but "false" clusters in the other direction). Higher RCS levels correspond to more positively-skewed cluster-size distributions with a large number of small clusters. Small

clusters, in turn, are harder to detect through periodic sampling studies, and there is a greater chance of misclassifying single-sampled members of these clusters toward a unique strain.

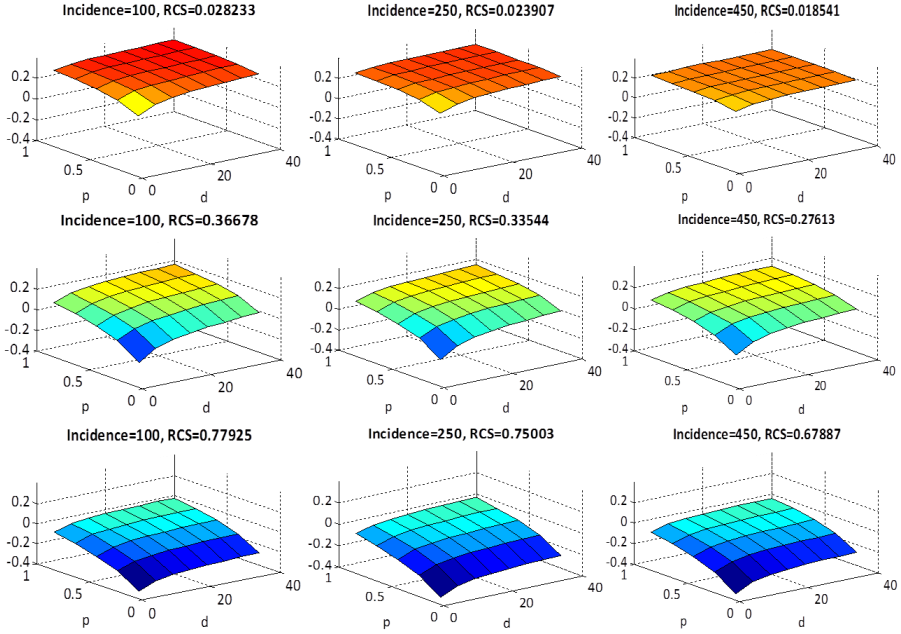

Figure 6: Comparison of Estimation Error Across Experiments.

Increasing the incidence level in each row (at fixed migration rates) causes a slight decrease in the estimate of recent transmission. As in the right-hand panel of Figure 4, the effects reflect the fact that higher incidence settings are characterized by a higher proportion of incidence due to primary infection and therefore also a lower RCS value. The increase in the estimated proportion of recent transmission is less than the increase in the true value, such that the overall estimation bias becomes more negative.

Finally, variation in the level of study coverage and sampling duration in each scenario suggest a similar effect on the estimation bias: lower sample representativeness (due to lower coverage and shorter duration) increases the underestimation of clustering level and moves the estimation error in a negative direction. Incomplete studies have a higher chance of missing clustered cases in a chain of transmission, underestimating the level of clustering and eventually underestimating the proportion of incidence due to recent transmission.

5 DISCUSSION

In this study, we propose a simulation-based approach to examine the role of various factors in population-based studies of TB molecular epidemiology for estimating the proportion of incidence due to recent transmission. We extend the literature by modeling the effect of various factors in relation to disease incidence, diversity of circulating strains, and study characteristics, and carrying wide sensitivity analysis of final estimation error for proportion of incidence due to recent transmission.

Our results agree with previous findings on the role of sampling incompleteness. Molecular analyses of samples acquired with low coverage or short duration reveal lower estimates of the underlying clustering level, with corresponding increases in the estimation bias of the proportion of incidence due to recent transmission. The distribution of cluster sizes, on the other hand, has a major effect on the final estimation bias. This distribution is influenced by the ongoing transmission dynamics (in relation to incidence) and strains mixing-pattern (in relation to migration rate) throughout the simulation period. Subsequently, higher RCS and lower incidence levels correspond to more positively-skewed cluster-size distributions, which in

turn reduce the overall estimate of the clustering level. As above, the magnitude of bias is greater for incomplete studies with a lower level of coverage and duration.

This study is an initial effort to improve estimates of TB recent transmission using a simulation-based approach in conjunction with molecular epidemiologic techniques. While several studies have described the shortcomings of the "*n*-1" method in restricted settings with regard to a single factor or a collection of factors in a specific scenario (Murray & Alland 2002), no successful attempt has yet been made to improve upon this method. Here we use a simulation-based approach to study the role of various factors in the presence of each other, and measure their effects on the precision of "*n*-1" results. These results may be useful in future work to develop monograms for the proportion of TB due to recent transmission that improve on "*n*-1" estimates. The ABS methodology provides a flexible and powerful platform for incorporating realistic assumptions and modeling transmission dynamics, and allows for explicit implementation of TB molecular studies at an individual- and population-based level. A main difficulty in such models, however, is establishing an appropriate process for retaining the strain diversity in the long run. This is particularly challenging in models of TB due to slow disease dynamics and the tendency to become latent, which also limit our understanding of TB transmission patterns. To address this problem, we demonstrate a migration-based procedure for modeling the mixing pattern of the study population with the "outside world." The "migration" rate does not correspond to actual migration in real populations, but rather is used to retain the diversity of strains over time. At high migration levels (e.g., 10% per year), the likelihood of strain exchange among recently infected individuals in the ELTB state increases, which subsequently limits the ability of the cluster-analysis method for identifying large transmission clusters that may be interrupted by "migration." Associations between the (population-based) migration rate and RCS depend on the underlying incidence rate, further complicating the calibration of RCS levels. Potential solutions for further analysis would be to limit "migration" to the LLTB state, or to replace the migration-based approach by an "external-infection" modeling logic in which individuals are subject to varying likelihoods of disease transmission from an external population.

Although the strain-type repositories generated by this model are not calibrated to represent realistic data sets, our results have implications for the practical interpretation of data in empirical studies. Specifically, the *"n-1"* cluster-analysis method, while known to have limitations, is still widely employed in epidemiological studies. In extensions of the current work, the data from these simulations will be used to construct a generalizable framework to estimate the amount of bias in the *"n-1"* method for any study of known duration, coverage, and incidence level. Such a model can then be used to adjust the current estimate of the amount of TB disease due to recent transmission from the cluster-analysis method, and therefore improve and refine current and future disease-control efforts.

6 REFERENCES

Andrews, J. R, F. Noubary, R. P. Walensky, R. Cerda, E. Losina, and C. R. Horsburgh. 2012. "Risk of Progression to Active Tuberculosis Following Reinfection with Mycobacterium Tuberculosis." *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America* 54 (6): 784–791.

Chevrel-Dellagi, D., A. M. E. L. Abderrahman, R. A. J. A. Haltiti, H. I. C. H. E. M. Koubaji, B. Gicquel, and K. Dellagi. 1993. "Large-Scale DNA Fingerprinting of Mycobacterium Tuberculosis Strains as a Tool for Epidemiological Studies of Tuberculosis." *Journal of Clinical Microbiology* 31 (9): 2446–2450.

Dowdy, D. W., S. Basu, and J. R. Andrews. "Is Passive Diagnosis Enough? The Impact of Subclinical Disease on Diagnostic Strategies for Tuberculosis." *American Journal of Respiratory and Critical Care Medicine* 187 (5): 543–551.

Dowdy, D. W., and R. E. Chaisson. 2009. "The Persistence of Tuberculosis in the Age of DOTS: Reassessing the Effect of Case Detection." *Bulletin of the World Health Organization* 87 (4): 296–304.

Kasaie, P., D. W. Dowdy, and W. D. Kelton. 2013. "An Agent-Based Simulation of a Tuberculosis Epidemic: Undestanding the Timing of Transmission." In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.H Kim, A. Tolk, R. Hill, and M.E. Kuhl, 2227–2238. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Moonan, P. K., S. Ghosh, J. E. Oeltmann, J. S. Kammerer, L. S. Cowan, and T. R. Navin. 2012. "Using Genotyping and Geospatial Scanning to Estimate Recent Mycobacterium Tuberculosis Transmission, United States." *Emerging Infectious Diseases* 18 (3): 458–465.

Murray, M. 2002a. "Sampling Bias in the Molecular Epidemiology of Tuberculosis." *Emerging Infectious Diseases* 8 (4): 363.

Murray, M. 2002b. "Determinants of Cluster Distribution in the Molecular Epidemiology of Tuberculosis." *Proceedings of the National Academy of Sciences* 99 (3): 1538–1543.

Murray, M., and D. Alland. 2002. "Methodological Problems in the Molecular Epidemiology of Tuberculosis." *American Journal of Epidemiology* 155 (6): 565–571.

Tiemersma, E. W., M. J. van der Werf, M. W. Borgdorff, B. G. Williams, and N. J. D. Nagelkerke. 2011. "Natural History of Tuberculosis: Duration and Fatality of Untreated Pulmonary Tuberculosis in HIV Negative Patients: A Systematic Review." *PLoS One* 6 (4): e17601.

Vynnycky, E. and P. E. Fine. 1997. "The Natural History of TB: The Implications of Age-Dependent Risks of Disease and Role of Reinfection." *Epidemiology and Infection* 119 (2): 183–201.

Vynnycky, E., N. Nagelkerke, M. W. Borgdorff, D. Van Soolingen, J. D. A. Van Embden, P. E. M. Fine, and others. 2001. "The Effect of Age and Study Duration on the Relationship Between'clustering'of DNA Fingerprint Patterns and the Proportion of Tuberculosis Disease Attributable to Recent Transmission." *Epidemiology and Infection* 126 (1): 43–62.

World Health Organization. 2013a. *Global Tuberculosis Report 2013*. http://www.who.int/tb/publications/global_report/en/.

World Health Organization. 2013b. *Global Health Observatory Data Repository-Life Expectancy: Life Tables by Country India*. http://apps.who.int/gho/data/view.main.60740?lang=en.

AUTHOR BIOGRAPHIES

PARASTU KASAIE is a postdoctoral fellow in the department of epidemiology at the Johns Hopkins Bloomberg School of Public Health. Her research interests are in simulation modeling and analysis of infectious disease epidemics and implications for policymaking. Her email address is pkasaie@jhu.edu.

DAVID W. DOWDY is the B. Frank and Kathleen Polk Assistant Professor of Epidemiology at the Johns Hopkins Bloomberg School of Public Health. His research interests include epidemiological investigation, economic evaluation, and mathematical modeling of tuberculosis and TB/HIV. He is particularly interested in the evaluation of diagnostic tests for TB and other infectious diseases and in development of "user-friendly" models for translation of findings to individuals without modeling expertise. Dr. Dowdy serves as a Steering Committee member of the TB Modeling and Analysis Consortium (funded by the Bill and Melinda Gates Foundation) and chair of the Modeling Subgroup of the TB Diagnostics Research Forum (co-funded by the Gates Foundation and U.S. National Institutes of Health). His email address is ddowdy@jhsph.edu.

W. DAVID KELTON is a Professor in the Department of Operations, Business Analytics, and Information Systems at the University of Cincinnati. His research interests and publications are in the probabilistic and statistical aspects of simulation, applications of simulation, and stochastic models. He is co-author of *Simio and Simulation: Modeling, Analysis, and Applications*, as well as *Simulation with Arena*, and was also coauthor of the first three editions of *Simulation Modeling and Analysis*. He was Editor-in-Chief of the *INFORMS Journal on Computing* from 2000-2007. He served as WSC Program Chair in 1987, General Chair in 1991, was on the WSC Board of Directors from 1991-1999, and is a Founding Trustee of the WSC Foundation. He is a Fellow of both INFORMS and IIE. His email address is david.kelton@uc.edu.