Model-based analysis of tuberculosis genotype clusters in the United States reveals high degree of heterogeneity in transmission, and state-level differences across California, Florida, New York, and Texas.

Sourya Shrestha[1, #], Kathryn Winglee[2], Andrew Hill[2], Tambi Shaw[3], Jonathan Smith[4], J. Steve Kammerer[2], Benjamin J. Silk[2], Suzanne Marks[2], David Dowdy [1]

[1]Johns Hopkins Bloomberg School of Public Health, Baltimore, US.

[2]Division of Tuberculosis Elimination, Centers for Disease Control and Prevention, Atlanta, US.

[3]California Department of Public Health, California, US.

[4]Yale University, New Haven CT, US.

[#] Corresponding author: Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, US; Email: sshres14@jh.edu

Short summary:

Tuberculosis transmission rates in the United States are low but highly heterogeneous; a small fraction of cases contribute substantially to overall transmission. Understanding the drivers of this heterogeneity could improve outbreak prevention reduce TB transmission.

ABSTRACT

Background: Reductions in tuberculosis (TB) transmission have been instrumental in lowering TB incidence in the United States. Sustaining and augmenting these reductions are key public health priorities.

Methods: We fit mechanistic transmission models to distributions of genotype clusters of TB cases reported to CDC during 2012–2016 in the United States and separately in California, Florida, New York, and Texas. Using these models, we estimated the mean number of secondary cases generated per infectious case (R_0) and individual-level heterogeneity in R_0 at state and national levels. We also assessed how different definitions of clustering and variation in case ascertainment affected these estimates.

Results: In clusters of genotypically linked TB cases occurring within a state over a 5-year period (reference scenario), the estimated R_0 was 0.29 (95% CI: 0.28–0.31) in the United States. Transmission was highly heterogeneous: 0.24% of simulated cases with individual R_0>10 generated 19% of all recent secondary transmissions. R_0 estimate was 0.16 (0.15–0.17) when a cluster was defined as cases occurring within the same county over a 3-year period. Transmission varied across states: estimated $R_0 s$ were 0.34 (0.3–0.4) in California, 0.28 (0.24–0.36) in Florida, 0.19 (0.15–0.27) in New York, and 0.38 (0.33–0.46) in Texas.

Conclusions: TB transmission in the United States is characterized by pronounced heterogeneity at the individual and state levels. Improving detection of transmission clusters through incorporation of whole-genome sequencing and identifying the drivers of this heterogeneity will be essential to reducing TB transmission in the United States and worldwide.

Introduction

Tuberculosis (TB) incidence in the United States fell by more than 70% between 1993 and 2017; reductions in transmission driven by progress in detecting and treating latent TB infection among persons recently exposed have been a key component of this decline.[1,2] Even though a minority of new TB cases are due to recent transmission,[3,4] extensive public health resources are required for their investigation and control. In the absence of timely TB control measures, these recent transmission outbreaks can grow, leading to large numbers of cases within the local community.[5] This is especially pertinent for vulnerable populations, which include racial and ethnic minorities, persons living in congregate settings (including correctional facilities and homeless shelters), and patients with medical comorbidities who are susceptible to TB and poor TB outcomes.[6–9] The risk of outbreaks expanding into larger populations and becoming endemic increases as they become larger or more frequent.[10] Understanding more about past outbreaks and responding with focused strategies can help prevent future outbreaks and mitigate these negative impacts.

Transmission of *Mycobacterium tuberculosis* (*Mtb)* is heterogeneous, driven by pathogen, host, environmental, and societal factors.[11] A better understanding of this heterogeneity can help improve TB control efforts, including outbreak prevention and response. Molecular characterization of *Mtb* isolates through genotyping

— which consists of matching isolates on the basis of both spacer oligonucleotide typing (spoligotype) and 24-locus mycobacterial interspersed repetitive unit-variable number of tandem repeats (MIRU-VNTR)—can identify TB cases that are clustered and thus presumptively related by recent transmission.[12] Genotyping has also been used to understand the evolution and spread of *Mtb.*[13,14] We use the U.S. distribution of genotypically clustered TB cases to characterize recent

transmission nationally and in the four states reporting over half of all TB cases: California, Florida, New York, and Texas.

Methods

Cluster distribution data

We used data from the U.S. National Tuberculosis Surveillance System (NTSS) and the National Tuberculosis Genotyping Service (NTGS) for TB cases reported from the 50 U.S. states and District of Columbia during 2012–2016 to infer the distribution of TB clusters in the United States and independently in California, Florida, New York, and Texas. Cases were defined as clustered if they (a) had matching spacer oligonucleotide typing (spoligotype) and 24-locus mycobacterial interspersed repetitive unit—variable number of tandem repeats (MIRU-VNTR) genotyping results, (b) were reported within specified geographic boundaries (i.e., same county or state), and (c) occurred during two time periods (i.e., 2012–2016 or 2014–2016). A cluster definition that included cases reported within a single state boundary and within 2012–2016 was defined as the reference scenario.

Branching Process Model

We used a branching process framework to describe recent transmission and cluster formation.[15,16] In this framework, the number of secondary cases resulting from a single case over a course of its infection occurs probabilistically and is given by the "offspring distribution" of the branching process model, Z (Table 1). The mean of this distribution is R_0, the reproductive number, equal to the average number of secondary cases resulting from a single case. We assume that this probability

distribution of secondary cases follows a Poisson distribution with parameter ν, i.e., $\boldsymbol{P}(Z = z) \sim \text{Poisson}(\nu)$.

Modeling individual-level heterogeneity

To incorporate individual-level heterogeneity in transmission, we varied the value of ν between individuals; here ν can also be interpreted as the individual-level reproductive number.[16] We compared four different models, with each encapsulating a different level of heterogeneity represented by the distribution of ν (Table 2). In all four models, the mean of both ν and the offspring distribution Z is R_0. Model I ("homogeneous" model) assumes that ν is constant ($\nu = R_0$) so that all individuals have the same infectious potential, and the number of secondary cases resulting from each case is Poisson-distributed. Model II ("SIR-type" model) assumes that the individual reproductive number is distributed exponentially, similar to assumptions in standard SIR-type compartmental models. Model III ("overdispersed" model) –a model previously used to capture heterogeneity in transmission of TB and other infectious diseases[14,16–18] –assumes a gamma distributed ν, resulting in a negative binomial distribution of secondary cases, with mean R_0 and shape parameter k. Finally, Model IV ("long-tailed" model) allows for even greater heterogeneity using a Poisson lognormal distribution; this modeling approach is often used to describe species abundance[19,20] and recently heterogeneity in TB transmission.[21] This model assumes that the individual reproductive numbers ν follow a lognormal distribution (i.e., $\exp(\nu) \sim N(\mu, \sigma^2)$), where μ and σ are the mean and standard deviation, respectively, of an underlying normal model. Larger values of σ^2 are indicative of increased heterogeneity.

Model comparison

We used a likelihood-based framework to evaluate and compare the fit of each model described above to the observed data. Using the likelihood function (described in detail in the supplementary materials, section S4), we calculated the maximum likelihood estimates (MLEs), the parameters that yield the highest likelihood, and corresponding 95% confidence regions/intervals. We compared models using Akaike information criterion, $AIC = 2r - 2\ln(\hat{\mathcal{L}})$, where r is the number of parameters in a model, and $\hat{\mathcal{L}}$ is the likelihood estimate.

Sensitivity of model inference to data censoring and importation

To assess the sensitivity of model inference to possible imperfections in data, we conducted a simulation study in which we considered two mechanisms by which observed data could differ from true clustering. First, we assumed underreporting of clustered TB cases due to factors such as cases not being reported in local jurisdictions, cases not being culture-confirmed (e.g., pediatric cases) or isolates not being genotyped, or cases being right- or left- censored over time. Second, we assumed over-ascertainment of clusters due to inclusion of imported cases of matching genotype (i.e., not due to local or recent transmission). We generated synthetic cluster distributions by simulating the branching process models under various assumptions about R_0 and individual-level heterogeneity (taken as true parameter values), which also incorporated imperfections in data described above. For each synthetic cluster distribution, we then applied the likelihood-based inference method to estimate both R_0 and individual level heterogeneity (estimated parameter values). By comparing true parameter values to their corresponding estimates, we inferred the sensitivity of each estimated parameter value to underreporting or over-ascertainment. (See section S5 and S6, and Figs S1-S4 for additional details.)

Results

Cluster distributions

Of 35,313 genotyped TB cases reported during 2012–2016 in the United States, 13,159 cases (37%) were clustered under the reference definition, of having the same genotyping result as at least one other case that was reported from the same state during the same period (Fig 1A). The remaining 22,154 (63%) unclustered cases did not have another same-state case with the same genotype during that timeframe. Among clustered cases, 31% occurred in large (≥10 cases) clusters; the largest cluster included 148 cases. When we restricted the definition of clustering to cases reported within county boundaries and occurring within a 3-year period (2014–2016), 21% of cases were clustered, and 15% of clustered cases were in clusters of ≥10 cases. The largest cluster contained 65 cases (Fig 1B).

Model comparison

Of the four models considered, Model IV ("long-tailed" model), which assumed the highest level of individual-level heterogeneity, provided the best statistical fit. The MLE under Model IV was statistically >1,000 times more likely to explain the data than Models I–III (Table 2). Much of this improved fit reflected a better ability to represent clusters of large size (i.e., the long tail), which occurred more frequently than predicted under Models I–III (Fig 2). This result was found consistently, regardless of cluster period or if clusters were defined within counties or states (see Figs S5–S8). Comparison of model fits across four states (Section S9, and Figs S11 and S12) also support this conclusion.

Individual-level heterogeneity in transmission

The estimated distribution of the individual reproductive number revealed substantial individual-level heterogeneity (Fig 3). For 95% of TB cases, the individual reproductive number was estimated to be less than one; these cases generated only 38% of secondary cases. Only 5% of TB cases were estimated to have individual reproductive number of 1 or greater, and these individuals generated 62% of secondary cases. Of note, 0.24% of cases were estimated to have a reproductive number larger than 10 and were projected to generate 19% of secondary cases.

Cluster ascertainment criteria and inference

The estimated value of R_0 ($\widehat{R_0}$) varied by geographic area and the timeframe used to define clustering. For example, $\widehat{R_0}$ in the reference scenario (state boundaries, and five-year timeframe between 2012 and 2016) was 0.29 (95% confidence interval [CI]: 0.28–0.31; Fig 4, hatched orange region); this estimate fell to 0.25 (95%CI: 0.24–0.27; Fig 4, orange region) when using a 3-year timeframe between 2014 and 2016, 0.19 (95%CI: 0.18–0.2; Fig 4, hatched green region) when using county boundaries but a 5-year window, and 0.16 (95%CI: 0.15–0.17; a 45% reduction; Fig 4, green region) when using both 3-year timeframe and county boundaries (See Table S1). Estimates of heterogeneity in transmission were less sensitive to the choice of cluster definition. For example, the estimated percentage of secondary cases originating from cases with R_0>10 fell between 16% and 19% regardless of cluster definition (Fig S9).

State-level variation

The cluster distribution of TB cases and the corresponding estimates of individual-level reproductive numbers varied considerably at the state level. For example, the proportion of clusters with ≥10 cases was nearly 8-fold larger in Texas compared to New York (Fig 5, blue line compared to red line). Consequently, the estimated mean individual-level reproductive numbers varied by a factor of two:

0.19 (95% CI: 0.15–0.27) in New York, 0.28 (95% CI: 0.24–0.36) in Florida, 0.34 (95% CI: 0.3–0.4) in California, and 0.38 (95% CI: 0.33–0.46) in Texas (See Table S2). There were substantial differences in estimated degree of transmission. For example, the contribution of individuals with R_0>10 to the total secondary cases varied from 9.5% (from 0.13% of individuals) in Florida to 20% (from 0.3% individuals) in California (Fig S10).

Sensitivity of model inference to data censoring and importation

Under- and over-ascertainment of clusters had a predictable effect on the inference of R_0. R_0 was underestimated (bottom left quadrant in Fig S3A) when cases were underreported and overestimated (top right quadrant in Fig S3A) when cases were over-ascertained. In both instances, the degree of under- or overestimation was linearly associated with the level of underreporting or over-ascertainment. Estimates of individual-level heterogeneity in transmission (σ) were unaffected by underreporting and were only slightly underestimated when the observed clusters included over-ascertainment of imported cases (Fig S3B).

Discussion

This model-based analysis of genotype-clustered TB cases in the United States revealed that there is substantial heterogeneity in transmission. We estimated that 95% of individual cases transmit to less than one secondary case each and contribute to only 38% of overall secondary transmission. By contrast, 0.24% of cases were estimated to transmit to 10 or more secondary cases, resulting in 19% of all secondary cases. This degree of heterogeneity is larger than described with prior models (i.e., negative binomial distribution)[16,22], but is remarkably consistent with data and prior analysis from the United Kingdom and the Netherlands[21] (Section S10, Figs S13 and S14). Taken together, these results suggest that heterogeneity in TB transmission (in low-burden settings) may be larger than

previously thought and may be driven by common underlying processes (e.g., propensity for transmission in outbreak-prone settings).

The characteristics of *Mtb* transmission varied across states. For example, even though TB incidence is similar in Texas and New York, the estimated R_0 in Texas was twice as high, suggesting that more cases in Texas reflect recent transmission, whereas more cases in New York may represent reactivation of latent infection or importation. These findings are consistent with estimates of recent transmission from the U.S. Centers for Disease Control and Prevention (CDC)[23] and from a previously published transmission model.[24] This may be reflecting differences in demography, immigration patterns, mobility, population density, relative sizes of key populations (e.g., people experiencing homelessness or incarceration), variation in *Mtb* strain pathogenicity and virulence, and societal context.[25] Also important to consider are differences in the size of jurisdictions, particularly counties, which can vary considerably between states: outbreaks are unlikely to be confined to small and/or highly interconnected counties (e.g., in cities) but may be contained in larger counties with more self-contained populations.

Conventional genotyping has known limitations that can lead to underestimation or overestimation of clustering. Underestimation may occur if true transmission-linked cases are not detected (e.g., individuals move out of a jurisdiction, or are reported elsewhere), do not have a specimen culture showing *Mtb*, or do not have an *Mtb* specimen that was genotyped (e.g, technical challenges associated with particular loci[26]). In the United States, most cases are likely diagnosed, and more than 95% of cultured cases have genotyped specimens,[27] making the choice of geographic area and period for the definition of clustering important. Our estimates suggest that defining clusters at the level of the county rather than the state could reduce estimates of transmission within clusters by

over one-third. Some left and right censoring in observed clusters may also occur over time. For example, cases may be missed towards the beginning of an outbreak (i.e., not included in the data) or towards the end (i.e., not yet diagnosed). Such under- or over-ascertainment might cause estimates of recent transmission (i.e., R_0) to fall or rise proportionally. The choice of time periods that we examined seemed to have less impact on estimates of transmission; none of the mechanisms mentioned above substantially affected estimates of individual-level heterogeneity in transmission, which remained high in all of our sensitivity analyses. When choosing the appropriate administrative level at which to define clusters, it is important to additionally consider the geographic size and population of administrative units, their interconnectedness, the relative value of a sensitive versus specific definition, and the level at which any response could be organized.

Conventional genotyping methods may also overestimate clustering by falsely attributing transmission links to cases that share common ancestry but are not related by recent transmission. TB often has a decades-long latency period, genotyping cannot be performed during this latent period, and molecular changes occur slowly; these factors can limit the use of conventional genotyping to estimate transmission. For example, cases resulting from a commonly circulating (endemic) strain might reactivate at similar times and thus could share a genotype but not reflect recent transmission events. Genotyping clusters defined by 24-loci MIRU-VNTR could encompass transmission events up to three decades in the past.[28] Studies that have compared genotyping with whole-genome sequencing (WGS), considered a gold standard, have found that only about 50%–60% of genotyped clusters could be confirmed by WGS.[29,30] This discrepancy tended to be much more pronounced for clusters consisting of nonnative individuals, suggesting that importation (i.e., transmission that occurred in distant past or outside of the country) can also contribute to overestimation.[26,29] Furthermore, accurate clustering by conventional genotyping varies by *Mtb* strain, with some genotypes and lineages having higher rates of internal diversity by WGS, meaning

they are less likely to be related by recent transmission.[31,32] These genotypes are also often concentrated in specific geographies, so the amount of inaccurate clustering can vary by state. As a result, varying imprecision in clustering could account for some of the variation in our estimates of heterogeneity of transmission between states.

Recent and local transmission can be corroborated through identification of epidemiologic links.[33] However, conducting epidemiologic investigations is challenging in populations with large numbers of clustered cases, especially if transmission occurs in settings and venues where identifying contacts may be infeasible (e.g., populations in which routine investigations are challenging, such as those experiencing homelessness and substance abuse).[34,35] Novel methods that incorporate transmission and epidemiological dynamics, contact patterns and network structures, genetic diversity, and evolutionary dynamics of *Mtb* may improve our ability to infer transmission by integrating molecular and genomic data.[18,36–41]

The high degree of heterogeneity in individual reproductive number estimated here might not only reflect individual-level factors, but environmental conditions, societal and healthcare provider-related factors that individuals experience. Communities or populations in which background *Mtb* infection rates are higher, comorbidities and risk factors are more prevalent, living conditions are more crowded and less ventilated (e.g., homeless shelters, correctional facilities), or more barriers to accessing health care exist are likely to experience higher rates of transmission and encounter challenges in the detection and treatment of latent TB infection among persons recently exposed to TB to prevent progression to infectious TB, resulting in larger and more frequent TB clusters.[42,43] Better characterization of larger clusters, for example by incorporating data on settings of possible transmission, demography, health equity measures, geography, and other risk factors,[44] might help

better identify factors that drive high transmission. In our analysis, we did not account for variability in infectiousness among clustered cases (e.g., sputum smear positivity and grade, radiographic evidence of cavitary disease among cases with pulmonary TB disease). Linkage of such clinical data with WGS and phylogenetic analyses of TB cases in large clusters might provide insight into the frequency and mechanism of high transmission, thus improving resource allocation by excluding probably unrelated cases from outbreak investigations. These insights could also help TB programs proactively identify (or potentially predict) the cases and contacts at highest risk of resulting in secondary cases and high transmission events so that further transmission can be prevented through intensified, focused interventions to ensure complete identification, evaluation, and treatment of recently infected contacts. Further work to identify if some strains of *Mtb* are associated with increased transmission due to pathogenicity or virulence factors could also help TB programs prioritize certain cases and contacts for follow up to prevent further transmission.

In conclusion, this model-based analysis of molecular surveillance data in the United States suggests that although the overall rate of recent TB transmission is generally low, a small fraction of TB cases probably plays an important role in driving transmission at the population level. Understanding the drivers of this heterogeneity—by identifying populations, settings, and activities that are more frequently associated with large outbreaks—could improve outbreak prevention and response (through early and accurate detection of large clusters), reduce TB transmission and improve TB-related resource allocation in the United States and more broadly.

NOTES

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention, or the views or opinions of the California Department of Public Health or the California Health and Human Services Agency.

Funding

This work was supported by the US Centers for Disease Control and Prevention, National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention Epidemiologic and Economic Modeling Agreement (grant number 5U38PS004646).

None of the authors has any potential conflicts to disclose.

References

1. Schwartz NG, Price SF, Pratt RH, Langer, AJ. Tuberculosis—United States, 2019. *MMWR Morb Mortal Wkly Rep*. 2020;69.

2. Armstrong LR, Winston CA, Stewart B, Tsang CA, Langer AJ, Navin TR. Changes in tuberculosis epidemiology, United States, 1993–2017. *Int J Tuberc Lung Dis*. 2019;23(7):797-804. doi:doi:10.5588/ijtld.18.0757

3. France AM, Grant J, Kammerer JS, Navin TR. A Field-Validated Approach Using Surveillance and Genotyping Data to Estimate Tuberculosis Attributable to Recent Transmission in the United States. *Am J Epidemiol*. 2015;182(9):799-807. doi:10.1093/aje/kwv121

4. Yuen CM, Kammerer JS, Marks K, Navin TR, France AM. Recent Transmission of Tuberculosis — United States, 2011–2014. *PLoS One*. 2016;11(4):1-13. doi:10.1371/journal.pone.0153728

5. Mindra G, Wortham JM, Haddad MB, Powell KM. Tuberculosis Outbreaks in the United States, 2009-2015. *Public Health Rep*. 2017;132(2):157-163. doi:10.1177/0033354916688270

6. Bamrah S, Yelk Woodruff RS, Powell K, Ghosh S, Kammerer JS, Haddad MB. Tuberculosis among the homeless, United States, 1994-2010. *Int J Tuberc Lung Dis*. 2013;17(11):1414-1419. doi:doi:10.5588/ijtld.13.0270

7. Baussano I, Williams BG, Nunn P, Beggiato M, Fedeli U, Scano F. Tuberculosis Incidence in Prisons: A Systematic Review. *PLOS Med*. 2010;7(12):1-10. doi:10.1371/journal.pmed.1000381

8. Valway SE, Greifinger RB, Papania M, et al. Multidrug-Resistant Tuberculosis in the New York State Prison System, 1990–1991. *J Infect Dis*. 1994;170(1):151-156. doi:10.1093/infdis/170.1.151

9. Warren JL, Grandjean L, Moore DAJ, et al. Investigating spillover of multidrug-resistant tuberculosis from a prison: a spatial and molecular epidemiological analysis. *BMC Med*. 2018;16(1):122. doi:10.1186/s12916-018-1111-x

10. Mathema B, Andrews JR, Cohen T, et al. Drivers of Tuberculosis Transmission. *J Infect Dis*. 2017;216(suppl_6):S644-S653. doi:10.1093/infdis/jix354

11. Trauer JM, Dodd PJ, Gomes MGM, et al. The Importance of Heterogeneity to the Epidemiology of Tuberculosis. *Clin Infect Dis*. 2018;69(1):159-166. doi:10.1093/cid/ciy938

12. Kammerer JS, Shang N, Althomsons SP, Haddad MB, Grant J, Navin TR. Using statistical methods and genotyping to detect tuberculosis outbreaks. *Int J Health Geogr*. 2013;12(1):15. doi:10.1186/1476-072X-12-15

13. van Soolingen D, Borgdorff MW, de Haas PEW, et al. Molecular Epidemiology of Tuberculosis in the Netherlands: A Nationwide Study from 1993 through 1997. *J Infect Dis*. 1999;180(3):726-736. doi:10.1086/314930

14. Ypma RJF, 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(3):395-400.

15. Farrington CP, Kanaan MN, Gay NJ. Branching process models for surveillance of infectious diseases controlled by mass vaccination. *Biostatistics*. 2003;4(2):279-295. doi:10.1093/biostatistics/4.2.279

16. Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. *Nature*. 2005;438(7066):355-359. doi:10.1038/nature04153

17. Blumberg S, Lloyd-Smith JO. Comparing methods for estimating R0 from the size distribution of subcritical transmission chains. *Epidemics*. 2013;5(3):131-145. doi:https://doi.org/10.1016/j.epidem.2013.05.002

18. 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(1):5382. https://doi.org/10.1038/s41598-018-23797-2.

19. Bulmer MG. On Fitting the Poisson Lognormal Distribution to Species-Abundance Data. *Biometrics*. 1974;30(1):101-110. http://www.jstor.org/stable/2529621.

20. Izsák R. Maximum likelihood fitting of the Poisson lognormal distribution. *Environ Ecol Stat*. 2008;15(2):143-156. doi:10.1007/s10651-007-0044-x

21. 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(3):1-14. doi:10.1371/journal.pcbi.1007687

22. Ypma RJF, van Ballegooijen WM, Wallinga J. Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. *Genetics*. 2013;195(3):1055-1062. doi:10.1534/genetics.113.154856

23. Centers for Disease Control and Prevention; National Center for HIV/AIDS, Viral Hepatitis, STD and T prevention. *2017 State and City Tuberculosis Report*.; 2018.

24. Shrestha S, Hill AN, Marks SM, Dowdy DW. Comparing Drivers and Dynamics of Tuberculosis (TB) in California, Florida, New York and Texas. *Am J Respir Crit Care Med*. 2017;196:1050-1059. doi:10.1164/rccm.201702-0377OC

25. Cherng ST, Shrestha S, Reynolds S, et al. Tuberculosis Incidence Among Populations at High Risk in California, Florida, New York, and Texas, 2011--2015. *Am J Public Health*. 2018;108(S4):S311--S314.

26. Teeter LD, Kammerer JS, Ghosh S, et al. Evaluation of 24-locus MIRU-VNTR genotyping in Mycobacterium tuberculosis cluster investigations in four jurisdictions in the United States, 2006–2010. *Tuberculosis*. 2017;106:9-15. doi:https://doi.org/10.1016/j.tube.2017.05.003

27. National Tuberculosis Surveillance System. Tuberculosis in the United States, 1993-2018. 2019. https://www.cdc.gov/tb/statistics/surv/surv2018/pdf/2018-surveillance-Report-Slideset.pdf.

28. Meehan CJ, Moris P, Kohl TA, et al. The relationship between transmission time and clustering methods in Mycobacterium tuberculosis epidemiology. *EBioMedicine*. 2018;37:410-416. doi:https://doi.org/10.1016/j.ebiom.2018.10.013

29. Stucki D, Ballif M, Egger M, et al. Standard Genotyping Overestimates Transmission of Mycobacterium tuberculosis among Immigrants in a Low-Incidence Country. K. C. C, ed. *J Clin Microbiol*. 2016;54(7):1862 LP - 1870. doi:10.1128/JCM.00126-16

30. Jajou R, Neeling A de, Hunen R van, et al. Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study. *PLoS One*. 2018;13(4):1-11. doi:10.1371/journal.pone.0195413

31. Wyllie DH, Davidson JA, Grace Smith E, et al. A Quantitative Evaluation of MIRU-VNTR Typing Against Whole-Genome Sequencing for Identifying Mycobacterium tuberculosis Transmission: A Prospective Observational Cohort Study. *EBioMedicine*. 2018;34:122-130. doi:https://doi.org/10.1016/j.ebiom.2018.07.019

32. Koster KJ, Largen A, Foster JT, et al. Genomic sequencing is required for identification of tuberculosis transmission in Hawaii. *BMC Infect Dis*. 2018;18(1):608. doi:10.1186/s12879-018-3502-1

33. Guthrie JL, Strudwick L, Roberts B, et al. Comparison of routine field epidemiology and whole genome sequencing to identify tuberculosis transmission in a remote setting. *Epidemiol Infect*. 2020;148:e15. doi:10.1017/S0950268820000072

34. Verver S, Warren RM, Munch Z, et al. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. *Lancet*. 2004;363(9404):212-214. doi:http://dx.doi.org/10.1016/S0140-6736(03)15332-9

35. Glynn JR, Guerra-Assunção JA, Houben RMGJ, et al. Whole Genome Sequencing Shows a Low Proportion of Tuberculosis Disease Is Attributable to Known Close Contacts in Rural Malawi. *PLoS One*. 2015;10(7):1-12. doi:10.1371/journal.pone.0132840

36. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. *Mol Biol Evol*. 2019;36(3):587-603. doi:10.1093/molbev/msy242

37. Hatherell H-A, Didelot X, Pollock SL, et al. Declaring a tuberculosis outbreak over with genomic epidemiology. *Microb Genomics*. 2016;2(5). doi:https://doi.org/10.1099/mgen.0.000060

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

doi:10.1093/molbev/msw275

39. Xu Y, Cancino-Muñoz I, Torres-Puente M, et al. High-resolution mapping of tuberculosis transmission: Whole genome sequencing and phylogenetic modelling of a cohort from Valencia Region, Spain. *PLOS Med*. 2019;16(10):1-20. doi:10.1371/journal.pmed.1002961

40. Nelson KN, Gandhi NR, Mathema B, et al. Modeling Missing Cases and Transmission Links in Networks of Extensively Drug-Resistant Tuberculosis in KwaZulu-Natal, South Africa. *Am J Epidemiol*. 2020;189(7):735-745. doi:10.1093/aje/kwaa028

41. Ma Y, Jenkins HE, Sebastiani P, et al. Using Cure Models to Estimate the Serial Interval of Tuberculosis With Limited Follow-up. *Am J Epidemiol*. 2020;189(11):1421-1426. doi:10.1093/aje/kwaa090

42. Moonan PK, Ghosh S, Oeltmann JE, Kammerer JS, Cowan LS, Navin TR. Using genotyping and geospatial scanning to estimate recent Mycobacterium tuberculosis transmission, United States. *Emerg Infect Dis*. 2012;18(3):458-465.

43. Haddad MB, Mitruka K, Oeltmann JE, Johns EB, Navin TR. Characteristics of Tuberculosis Cases that Started Outbreaks in the United States, 2002-2011. *Emerg Infect Dis J*. 2015;21(3):508. http://wwwnc.cdc.gov/eid/article/21/3/14-1475.

44. Division of Tuberculosis Elimination. Reported Tuberculosis in the United States, 2019. 2020. https://www.cdc.gov/tb/statistics/reports/2019/table59.htm.

Notations and symbols	Descriptions	Underlying assumptions
$\boldsymbol{R_0}$	Reproductive number, or average number of secondary cases resulting from a single case.	Theoretical concept.
$\boldsymbol{\nu}$	Individual reproductive number, expected number of secondary cases resulting from each individual	Assumed to vary based on the underlying models. See Table 2.
$\widehat{\boldsymbol{R_0}}$	Estimated reproductive number, based on maximum likelihood estimate.	The estimates are aimed to capture cases resulting from recent transmission (and exclude cases resulting from reactivation that occur at longer time scales). The estimates are based on genotyped cluster data, and subject to limitations of the clustering method (including missing cases, cases not cultured or genotyped, over-ascertainment of clustering).
$\boldsymbol{Z}$	Offspring distribution of a branching process, that describes the probability distribution of the number of secondary cases resulting from a single case	Varies based on the underlying models. See Table 2.

Table 1. Table of notations and symbols used in the paper along with detailed descriptions and underlying assumptions.

Models	Model description	Underlying distribution of individual reproductive number, ν; the resulting distribution of secondary cases, $\boldsymbol{Z}$; variance of $\boldsymbol{Z}$	Maximum likelihood estimate, MLE, log scaled (difference in log likelihood units relative to the highest estimate)	Relative likelihood compared to the best model **
Model I: Homogeneous model*	Assumes no individual-level heterogeneity, i.e., all individuals have the reproductive number.	ν is constant; $Z \sim \text{Poisson}(R_0)$; R_0	−16,787.68 (−1,450.19)	<1/1000
Model II: SIR-type model*	Reflecting assumption in standard SIR-type compartmental models, assumes exponentially distributed individual reproductive numbers.	ν is exponentially distributed; $Z \sim \text{geometric}(R_0)$; $R_0(1 + R_0)$	−17,804.98 (−2,468.19)	<1/1000
Model III: Overdispersed model	Assumes that the number of secondary cases from an individual are over dispersed, and the degree of overdispersion is estimated.	ν is gamma distributed; $Z \sim \text{negative binomial}(R_0, k)$ k is the dispersion parameter, smaller values relate to larger heterogeneity; $R_0(1 + \frac{R_0}{k})$	−15,507.78 (−170.99)	<1/1000
Model IV: Long-tailed model	Assumes that individual-level heterogeneity is lognormally distributed (allowing for even larger heterogeneity).	ν is lognormally distributed; $Z \sim \text{Poisson lognormal}(\mu, \sigma^2)$ μ, σ^2 are, respectively, mean variance of the underlying normal distribution; $R_0\,[1 + R_0\,(exp(\sigma^2) - 1)]$	**−15,336.79 (Ref)**	–

Table 2. Description of four models of individual-level heterogeneity, and comparison of their statistical fits to the reference data.

* Poisson and geometric models are specific instances of the negative binomial model: negative binomial model with dispersion parameter k→ ∞ is a Poisson model, and k=1 is a geometric model.

** Relative likelihood is given by the quantity $\exp((AIC_{min} - AIC)/2)$, where AIC is the Akaike information criterion, and AIC_{min} is the AIC score corresponding to the best-fit model, or the model with the lowest score.

FIGURE LEGENDS

Figure 1. Genotype cluster distribution of tuberculosis (TB) cases in the United States. Shown are the frequency of observed genotype TB clusters of various sizes in the United States based on (A) cases reported within a given state and occurring within a 5-year time period (2012 to 2016); and (B) cases reported within a given county and occurring within a 3-year time period (2014 to 2016). Genotypic clusters are defined as cases with matching spoligotype and 24-locus MIRU-VNTR occurring within the specified geographic boundary and during the specified time window. Both axes are plotted on a log scale.

Figure 2. Fitting branching process models to genotype cluster distributions of tuberculosis (TB) in the United States. We fit branching process models to the cluster distribution consisting of genotyped TB cases occurring within U.S. state boundaries over a 5-year time period (shown in Fig 1A). We considered four different model assumptions to describe the underlying individual-level heterogeneity. Model I (homogeneous) hypothesized that there was no individual-level heterogeneity except by stochastic chance alone (green), Model II (SIR-type) hypothesized that the individual reproductive number followed an exponential distribution (purple); Model III (overdispersed), a gamma distribution (yellow); and Model IV (long-tailed), a log-normal distribution (blue). Shown are (A) frequency distributions and (B) cumulative probability distributions corresponding to the best-fit models of each type (shown by colored dashed lines) against the data (shown in grey dots). Cluster size and frequency distributions are plotted on a log-scale, and the cumulative probability distribution is plotted on a logit scale.

Figure 3. Underlying individual-level heterogeneity of tuberculosis (TB) transmission. Shown is the probability density function corresponding to the best-fit Poisson lognormal model, describing the distribution of the individual reproductive number under the reference scenario (clustering based on genotyped cases reported within state boundary and occurring between 2012 and 2016). The solid vertical line shows the mean of the distribution (i.e., $\widehat{R_0}$), or the estimated average number of secondary transmission cases resulting from a single TB case.

Figure 4. Comparing model-based inferences under different definitions of tuberculosis clusters in the United States. We fit Poisson lognormal models to four separate cluster distributions, each using a different geographic boundary and time window for cluster ascertainment. Shown are the estimated mean reproductive number (i.e., $\widehat{R_0}$) and variance of the distribution of secondary cases, when the clusters were defined to include cases reported within (i) state boundaries and occurring

within 5-year time window (hatched orange); (ii) state boundaries and occurring within a 3-year time window (orange); (iii) county boundaries and occurring within a 5-year time window (hatched green); and (iv) county boundaries and occurring within a 3-year time window (green). The crosses indicate maximum likelihood estimates, and shaded areas indicate estimated 95% confidence intervals.

Figure 5. State-level heterogeneity in tuberculosis cluster distributions and transmission across four U.S. states: California, Florida, New York, and Texas between 2014-2016. (A) Colored circles show cluster distributions in California (green), Florida (violet), New York (blue), and Texas (orange) as cumulative probabilities (i.e., the probability of a cluster of given size or less). The colored lines show cumulative probabilities corresponding to the Poisson lognormal model with maximum likelihood estimates. (B-E) Shown are the estimated reproductive numbers $\widehat{R_0}s$ and variances (in the distribution of secondary cases) for California (B), Florida (C), New York (D), and Texas (E). The shaded colored region indicates the estimated 95% confidence region, and the cross in the middle indicates the maximum likelihood estimate.