PLOS ONE

RESEARCH ARTICLE

A Novel Tool Improves Existing Estimates of Recent Tuberculosis Transmission in Settings of Sparse Data Collection

Parastu Kasaie[1], Barun Mathema[2], W. David Kelton[3], Andrew S. Azman[1], Jeff Pennington[1], David W. Dowdy[1]*

1 Department of Epidemiology, Bloomberg School of Public Health, The Johns Hopkins University, Baltimore, MD, United States of America, **2** Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, United States of America, **3** Department of Operations, Business Analytics, and Information Systems, University of Cincinnati, Cincinnati, OH, United States of America

* ddowdy@jhsph.edu

Citation: Kasaie P, Mathema B, Kelton WD, Azman AS, Pennington J, Dowdy DW (2015) A Novel Tool Improves Existing Estimates of Recent Tuberculosis Transmission in Settings of Sparse Data Collection. PLoS ONE 10(12): e0144137. doi:10.1371/journal.pone.0144137

Editor: Clive M. Gray, University of Cape Town, SOUTH AFRICA

Received: August 3, 2015

Accepted: November 14, 2015

Published: December 17, 2015

Data Availability Statement: An original version of the model and all simulation classes has been uploaded on github and is accessible at https://github.com/pkasaie/TB_RecentTransmissionProportion.git.

Funding: This work was supported by B. Frank and Kathleen Polk Assistant Professorship in Epidemiology, Johns Hopkins Bloomberg School of Public Health [DWD, ASA, PK]. This research was funded in part by a 2015 developmental grant from the Johns Hopkins University Center for AIDS Research, an NIH funded program (P30AI094189), which is supported by the following NIH Co-Funding

Abstract

In any setting, a proportion of incident active tuberculosis (TB) reflects recent transmission ("recent transmission proportion"), whereas the remainder represents reactivation. Appropriately estimating the recent transmission proportion has important implications for local TB control, but existing approaches have known biases, especially where data are incomplete. We constructed a stochastic individual-based model of a TB epidemic and designed a set of simulations (derivation set) to develop two regression-based tools for estimating the recent transmission proportion from five inputs: underlying TB incidence, sampling coverage, study duration, clustered proportion of observed cases, and proportion of observed clusters in the sample. We tested these tools on a set of unrelated simulations (validation set), and compared their performance against that of the traditional '*n*-1' approach. In the validation set, the regression tools reduced the absolute estimation bias (difference between estimated and true recent transmission proportion) in the '*n*-1' technique by a median [interquartile range] of 60% [9%, 82%] and 69% [30%, 87%]. The bias in the '*n*-1' model was highly sensitive to underlying levels of study coverage and duration, and substantially underestimated the recent transmission proportion in settings of incomplete data coverage. By contrast, the regression models' performance was more consistent across different epidemiological settings and study characteristics. We provide one of these regression models as a user-friendly, web-based tool. Novel tools can improve our ability to estimate the recent TB transmission proportion from data that are observable (or estimable) by public health practitioners with limited available molecular data.

Introduction

Tuberculosis (TB) is unique among major infectious diseases in its ability to cause symptomatic and infectious (active) disease many years after transmission [1–3]. As a result, new cases

and Participating Institutes and Centers: NIAID, NCI, NICHD, NHLBI, NIDA, NIMH,NIA, FIC, NIGMS, NIDDK, and OAR. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Competing Interests: The authors have declared that no competing interests exist.

of active TB represent a mix of recent transmission and remote infection (reactivation) [4–6]. Understanding the proportion of TB incidence due to recent transmission versus reactivation–a quantity that we term the "recent transmission proportion"–has important implications for implementation of TB control strategies, as different strategies (for example, contact investigation, improved diagnosis of active disease) may have greater epidemiological impact in settings of ongoing transmission [7–11], whereas other strategies (for example, preventive therapy) may be more relevant for settings where most active TB represents reactivation [12–14].

The traditional approach for estimating the recent transmission proportion is based on analysis of DNA fingerprints [15,16]. TB cases that are linked by recent transmission events should have similar DNA fingerprints (forming "clusters"), whereas those that represent reactivation will generally differ. Such fingerprinting methods are becoming increasingly important in assessing recent transmission in settings of both low [17–22] and high TB burden [23–25] for purposes of public health planning. While the molecular methods used for fingerprinting vary in their discriminatory power, and thus the epidemiological relevance of identified clusters, novel technologies (e.g., whole-genome sequencing) continue to reduce the probability of clustering by chance [26,27], thereby improving our ability to discriminate between reactivation and recent transmission.

Unfortunately, technological advances in fingerprinting techniques have not been mirrored by analytical advances in estimating the recent transmission proportion [28],which is still frequently calculated using the simplistic assumption that, in each cluster, one case represents reactivation (the index) whereas all other cases represent recent transmission–an approach known as the '*n*-1' method [5]. The '*n*-1' method has several known biases [29–33] that limit its public health utility in settings where fingerprint data are incomplete or collected over a short time period, even when the fingerprinting technique itself is highly discriminatory. This is a broadly applicable problem in evaluating the dynamics of infectious diseases with a long transmission time scale (e.g., HIV, hepatitis C [34–36]), where molecular or genetic clustering data from population-based studies with limited coverage are often used to draw inference as to the proportion of transmission that occurs in a specific setting or on a given timescale. A less biased, user-friendly alternative for estimating the recent transmission proportion would therefore enable public health practitioners to maximize the utility of their molecular data. We therefore built a model of a generalized TB epidemic (Fig 1), and created a set of controlled simulation experiments ("*derivation set*") to develop an improved tool for estimating the recent transmission proportion from molecular fingerprint data. We compared the performance of this tool against the traditional '*n*-1' model, and validated the findings using an independent set of simulations ("*validation set*"). In order to illustrate the performance of this tool in a real setting, we also investigated its potential use on published data from Cape Town, South Africa [23].

Methods

We constructed a stochastic, individual-based simulation model of a TB epidemic that incorporates elements of TB natural history, TB epidemiology, and hypothetical molecular epidemiological studies (Fig 1). We modeled a hypothetical population of 100,000 individuals with homogenous mixing structure, calibrating birth and death rates to preserve the mean population size over time. TB natural history is modeled at an individual level in five states: uninfected, early latent, late latent, active disease, and recovered (Fig 1 and Table 1) [37]. A complete description of the model is provided in Section A of the S1 Appendices.

To capture the clustering dynamics of TB strains, we defined individual TB genotypes and modeled each transmission as resulting in an infection with a strain that, if isolated and fingerprinted, could be linked to the source case. Reinfection may occur, with an individual's

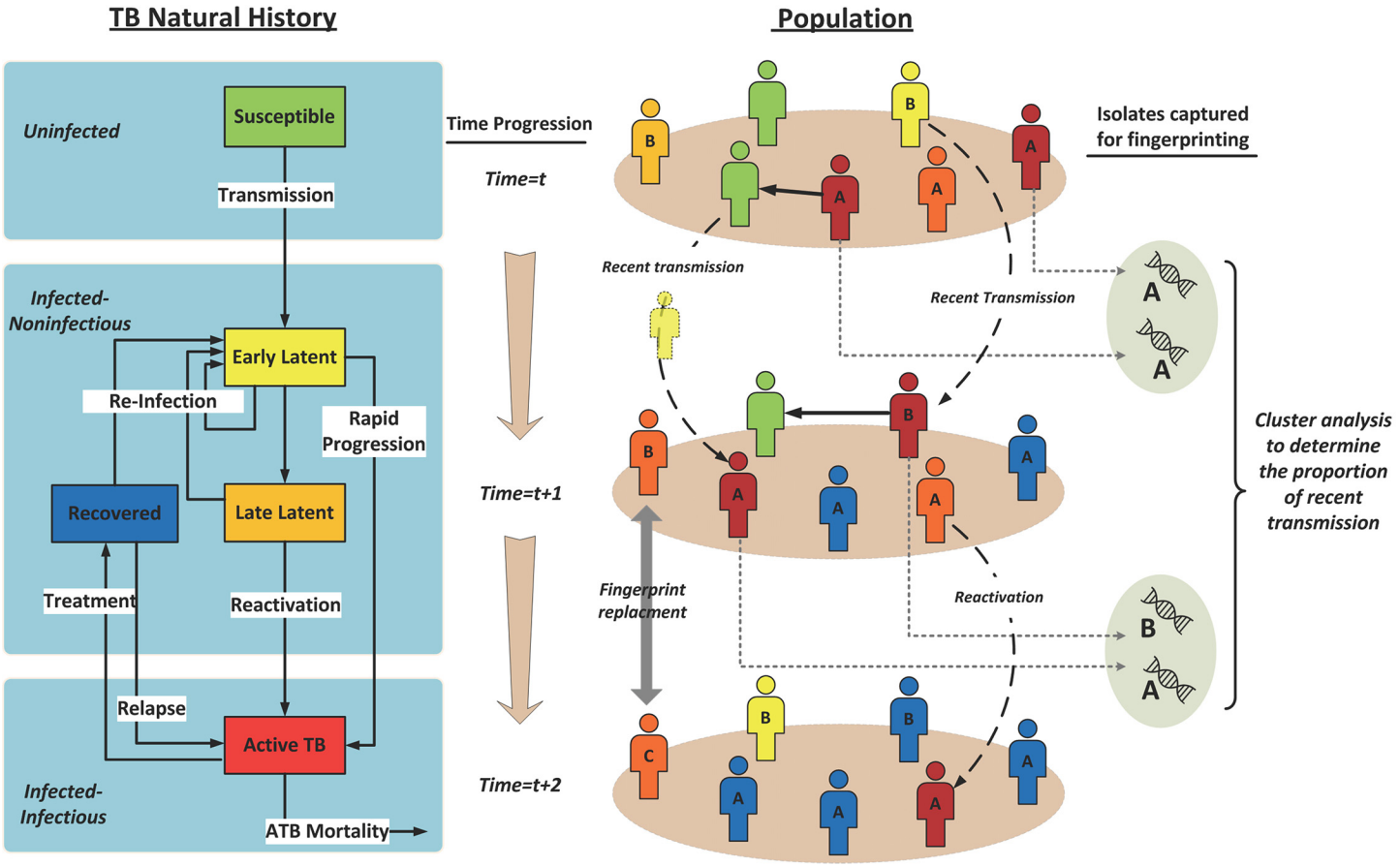

Fig 1. TB simulation overview. This figure illustrates our individual-based simulation model, following a hypothetical population for three consecutive (annual) time steps (Time = t, t+1, t+2). All infected individuals carry a single strain of TB (A, B, or C in this example). At each time step, three processes are modeled: 1. Transmission: upon successful contact, actively infected individuals can transmit the disease (marked by their strain type) to other people in the population. 2. Progression: other TB states are updated as shown in the left panel, including stabilization of latency, re-infection, diagnosis, and treatment, and relapse. Individuals who are diagnosed have their strain type recorded for analysis as they move from the active to the recovered state. 3. DNA fingerprint replacement: a random number of individuals in the late latency state are selected to carry new and unique fingerprints (strains), to maintain genetic diversity and account for processes such as mutation, migration, and infection from outside the population.

doi:10.1371/journal.pone.0144137.g001

simulated DNA fingerprint reflecting the most recent transmission event. Over time, the simulated closed population develops more strain homogeneity than observed in real populations where migration and bacterial polymorphisms introduce additional diversity. We thus instituted a "fingerprint replacement" process in which a random set of individuals with latent infection change mycobacterial genotypes each year [37], representing a combination of infection from individuals outside the population, immigration/emigration, and bacterial evolution [43,44]. To minimize potential bias of this approach (to maintain long-term diversity in each simulation, whereas clustering is measured in the short term), we performed this procedure only on individuals who were infected more than five years previously. The replacement rate was calibrated to observed levels of TB genetic diversity [28,45] (Section A.1. of the S1 Appendices).

Simulation experiments

After calibration, we developed a "derivation" set of simulations designed to cover setting-specific variables at regular intervals (for example, sampling TB incidence in intervals of 50 per

Table 1. Model parameters.

Parameter	Value / [Range]	Source/Description
Natural History:		
Cumulative proportion of TB infections that progress to active disease over five years	13.8%	[38]
Mortality rate from untreated active TB	0.12 per year	[39]
Rate of successful diagnosis and treatment of active TB	0.9 per year	Calibrated to provide prevalence/incidence ratio of 1.4, accounting for people on treatment [40]
Rate of relapse to active TB during the first two years after treatment	0.06 per year	Calibrated to provide 15% annual cumulative incidence among previously treated individuals [40]
Proportion of TB re-infections that successfully replace a latently established strain	50%	[41–42]
Setting-Specific:		
Replacement rate	[0.5–10]x10-3 per year	Derivation set samples the following six values: (0.5, 1, 2, 3, 4, 5, 7, 10) × 10–3
Incidence rate	[100–450] per 100,000 per year	Derivation set samples values in increments of 50 (e.g., 100, 150, 200, etc.); calibrated by changing the contact rate at each level of the replacement rate above.
Duration of time over which isolates are collected	[2–20] years	Derivation set samples values in increments of 2
Proportion of the population contributing isolates for fingerprinting	[20–100]%	Derivation set samples values in increments of 20%

doi:10.1371/journal.pone.0144137.t001

100,000/year), with natural history parameters held fixed at their best estimated value. Simulation scenarios were characterized by level of disease incidence as well as underlying proportion of incidence due to recent transmission (calibrated via contact rate and annual reactivation rate parameters). Limiting the scope of our study to settings of medium-to-high TB burden, we let the incidence level vary over the range of 100 to 450 cases per 100,000 per year [46], covering the parameter space using fixed intervals. Once TB natural history parameters were established, we then simulated different data collection exercises, in which a given proportion of individuals diagnosed with active TB (the sampling coverage) was fingerprinted over a specified period of time (the sampling duration). From each simulated "fingerprinting dataset", we calculated the corresponding number of clustered cases (C) and the number of clusters (N) that would be observed, assuming a fingerprinting technique of perfect discrimination. The traditional 'n-1' estimate of the recent transmission proportion can then be calculated as $(N\text{-}C)/SS$, where SS is the sample size (Section A.2 of the S1 Appendices). We compared this 'n-1' estimate with the actual (simulated) recent transmission proportion in each scenario to calculate the bias in the 'n-1' estimate. We defined the "estimation bias" as the absolute underestimation or overestimation of the true recent transmission proportion.

After calculating the 'n-1' estimate and its bias, we used the simulated data in the derivation set to develop an improved estimator of the recent transmission proportion via multiple linear regression incorporating five input variables: incidence, duration, coverage, ratio of clustered cases in the sample ($c = C/SS$), and ratio of observed clusters in the sample ($n = N/SS$). We evaluated both a simple linear model with these five covariates, as well as a more comprehensive model in which all potential multiplicative interaction terms were considered; models with even greater detail did not provide significant improvement in fit (data not shown). Full

regression equations are provided in Section C.1 of the S1 Appendices and as an online calculator (http://modeltb.org/recenttrans/).

To validate the performance of the regression models and study the sensitivity of results to variation of simulation parameters, we created a second set of simulations (the "validation set") in which both natural history and setting-specific parameters were randomly sampled from wide underlying distributions (Section B of the S1 Appendices). The parameter space was sampled using a Latin hypercube design to generate a sample of 1000 scenarios, of which 518 corresponded to a incidence level of 100 to 450 per 100,000/year (as used in original analysis). Data-collection exercises using varying levels of study coverage and duration were subsequently modeled in each scenario. In each simulation, we calculated the recent transmission proportion with both the traditional '*n*-1' method and the novel regression tools, and compared these to the "true" (simulated) value. We then calculated the bias in each estimation method as the difference between the estimated and the true value. We used partial rank correlation coefficients to evaluate associations between model input parameters and the resulting bias in each estimation method [47], and performed additional sensitivity analyses around the fingerprint replacement rate and in settings of very high incidence (sections D.2 and D.1 of the S1 Appendices).

Finally, to investigate the potential performance of suggested estimators in real settings, we investigated a representative case study from the literature of a genotyping study in Cape Town, South Africa, from 1993 to 1998 [23]. Calibrating the simulation model to the corresponding study settings, we compared the estimation bias of the '*n*-1' approach and regression-based models for the recent transmission proportion. Due to uncertainty in estimating the underlying TB incidence from available data on notifications, we considered two scenarios using a high case detection ratio (percentage of incident TB cases that are notified) of 90% similar to that assumed in the simulation derivation set (scenario 1), and a lower value equal to the current estimate of 62% for South Africa (scenario 2). Under these scenarios, we calibrated the replacement rate to produce similar measures of clustering as were observed in the original study (Section E of the S1 Appendices).

Results

Regression-based Tool Performance

In the derivation set of simulation experiments, which sampled key parameters (TB incidence and replacement rate) at regular intervals (Table 1), the median [interquartile range] absolute bias in the estimated recent transmission proportion was 7.8% [3.9%, 12.3%] with the '*n*-1' method (Fig 2- top graph). In other words, the median absolute difference between the recent transmission proportion as actually simulated versus as estimated by the '*n*-1' method was 8% (for example, 50% vs. 58%, or 75% vs. 83%). The simple regression model reduced this bias to 3.0% [1.4%, 4.6%], equivalent to a 65% [20%, 80%] reduction in the estimation bias of the '*n*-1' method, and the comprehensive regression model reduced the bias further to 1.5% [0.6%, 3.0%], or a 78% [58%, 94%] reduction. The statistical significance of improvements was further confirmed via A Wilcoxon signed-ranked test by comparing the absolute estimation bias in the "*n-1*" method with the regression models (p-value$<$ 0.001). When applied to the validation set, similar results were seen: median absolute estimation bias of 7.4% [3.6%, 12.2%] with the '*n*-1' method, 3.1% [1.5%, 4.9%] with the simple regression model (60% [9%, 82%] reduction in bias), and 2.3% [1.1%, 3.8%] with the comprehensive regression model (69% [30%, 87%] reduction in bias) (Fig 2-bottom graph).

As an outcome that might be relevant from a public health perspective, we measured the proportion of simulations in which the recent transmission proportion was over- or underestimated by 10% or more; 35% of simulations exceeded this threshold with the '*n*-1' method,

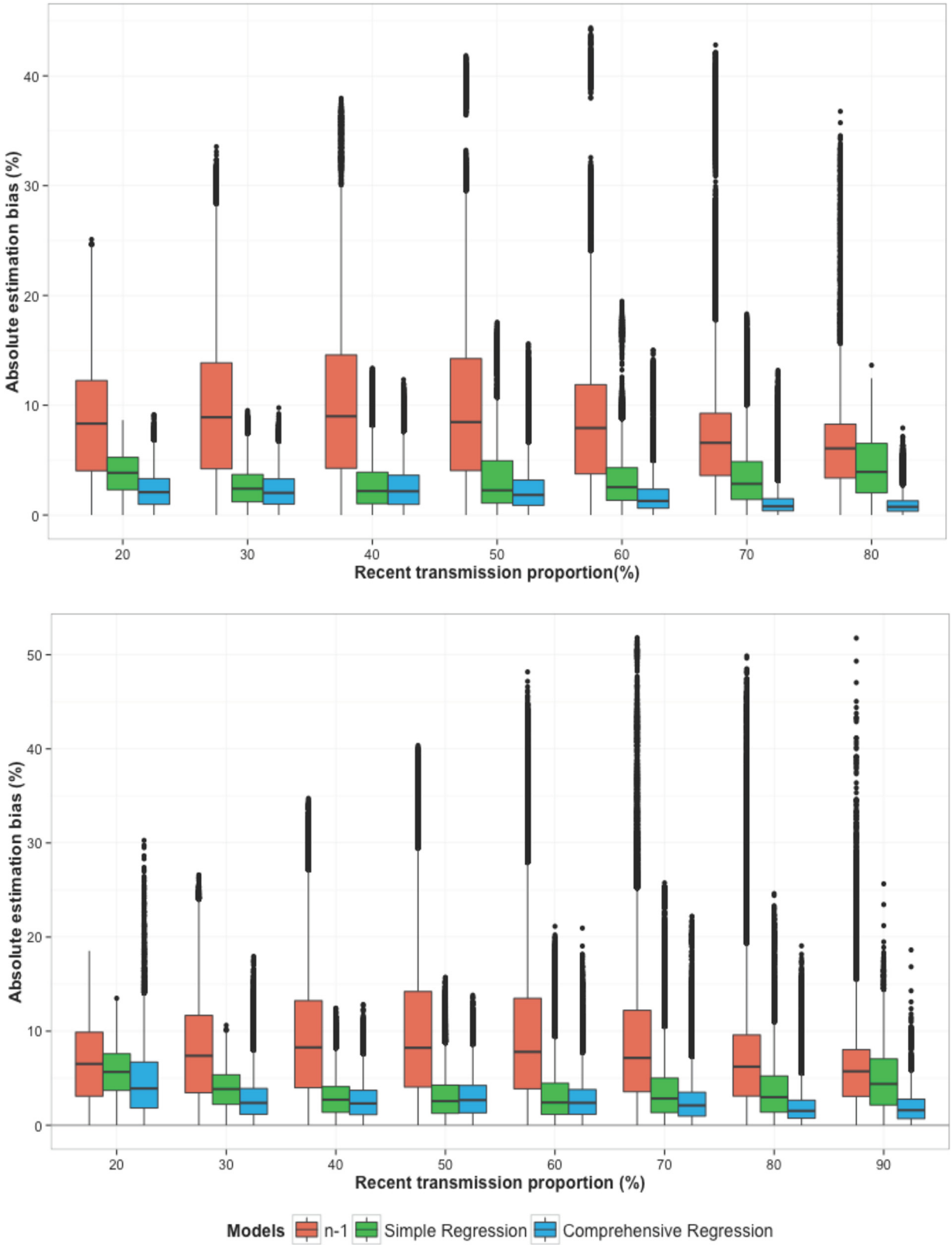
Absolute estimation bias (%)
0
10
20
30
40
20
30
40
50
60
70
80
Recent transmission proportion(%)
Absolute estimation bias (%)
0
10
20
30
40
50
20
30
40
50
60
70
80
90
Recent transmission proportion (%)
Models
n-1
Simple Regression
Comprehensive Regression

Fig 2. Absolute bias in estimates of the TB recent transmission proportion, comparing the 'n-1' method to novel regression-based tools in the derivation set (top) and validation set (bottom). The y-axis presents the absolute estimation bias [|estimated value – true value| × 100] in the proportion of incident active TB due to recent transmission ("recent transmission proportion"), and the x-axis denotes the recent transmission proportion (simulated) in each set of simulations. Estimates from the '*n-1*' method are shown in red, and those from the simple and comprehensive regression tools are shown in green and blue, respectively. Boxes show the interquartile range of values from all simulations, and "whiskers" show the 95% confidence intervals, such that narrower boxes correspond to more precise (reproducible) estimates. The 'n-1' model tends to provide less accurate and precise estimates of recent transmission proportion (wide red bars) across all settings as compared to the simple and comprehensive regression-based models.

doi:10.1371/journal.pone.0144137.g002

versus 2% with the simple regression model, and 1% with the comprehensive regression model (Fig 3, grey dotted line). Thus, the '*n-1*' method produced results with an absolute bias of more than 10% in over one-third of all simulated scenarios, whereas the regression tools resulted in such bias in less than one of every 50 simulations.

Moreover, the '*n*-1' method's accuracy in estimating the recent transmission proportion was highly sensitive to underlying levels of sampling coverage and duration (Fig 4). Specifically, fingerprinting studies with low coverage or short duration were more likely to generate

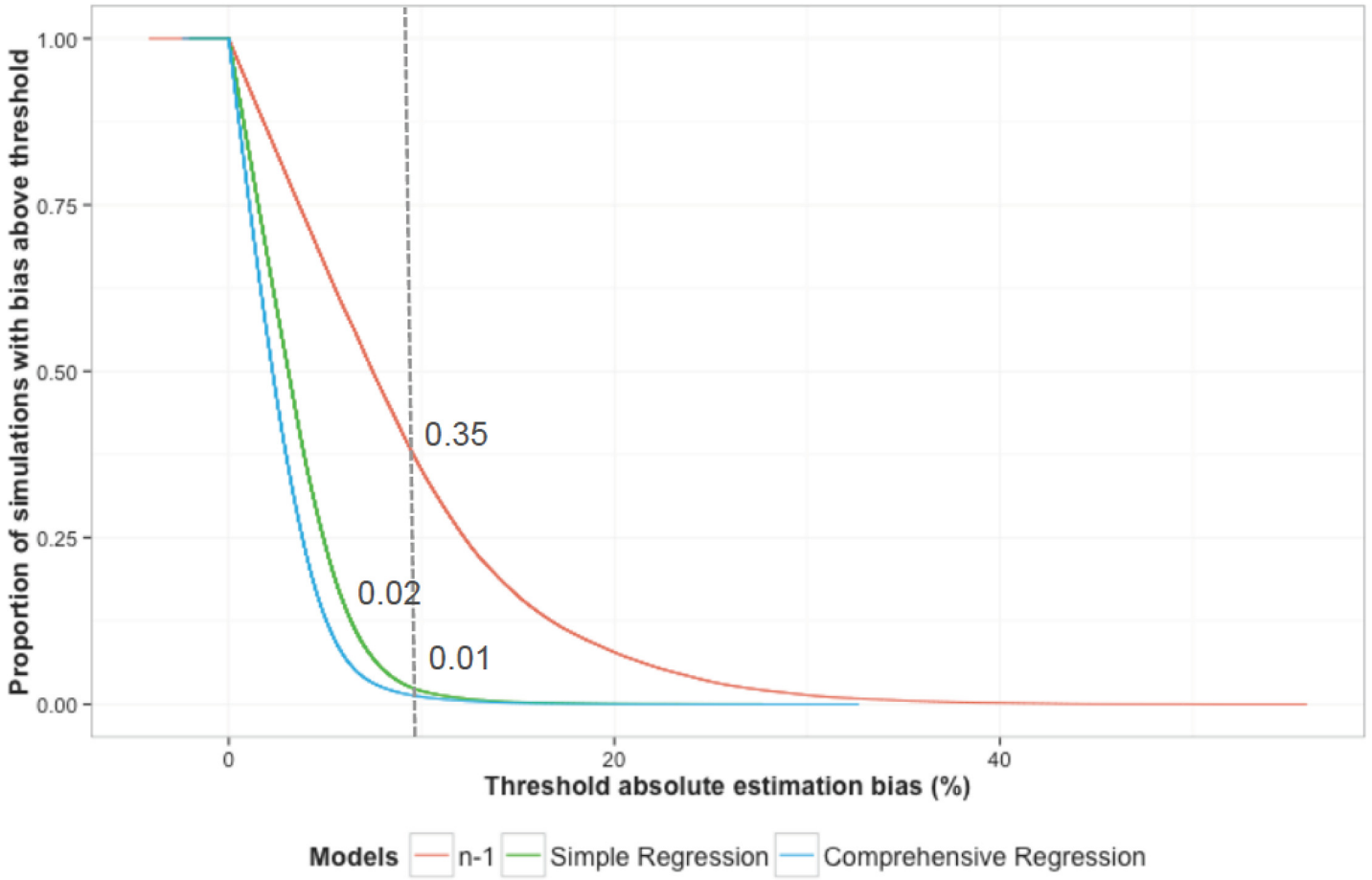

Fig 3. Absolute estimation bias in validation set, comparing the 'n-1' method to the regression-based models. The x-axis denotes the absolute level of estimation bias (abs[estimated value – true value]*100) in the proportion of incident active TB due to recent transmission ("recent transmission proportion") across the validation set of simulations. The y-axis denotes the cumulative proportion of simulations with estimation bias greater than the threshold shown on the x-axis. For example, the vertical dotted line shows the proportion of simulations under each method that resulted in an absolute estimation bias of >10%: the '*n-1*' method resulted in 10% or greater estimation bias in 35% of all simulations (red line), compared to 2% of all simulations with the simple regression model (green) and 1% of all simulations with the comprehensive regression model (blue).

doi:10.1371/journal.pone.0144137.g003

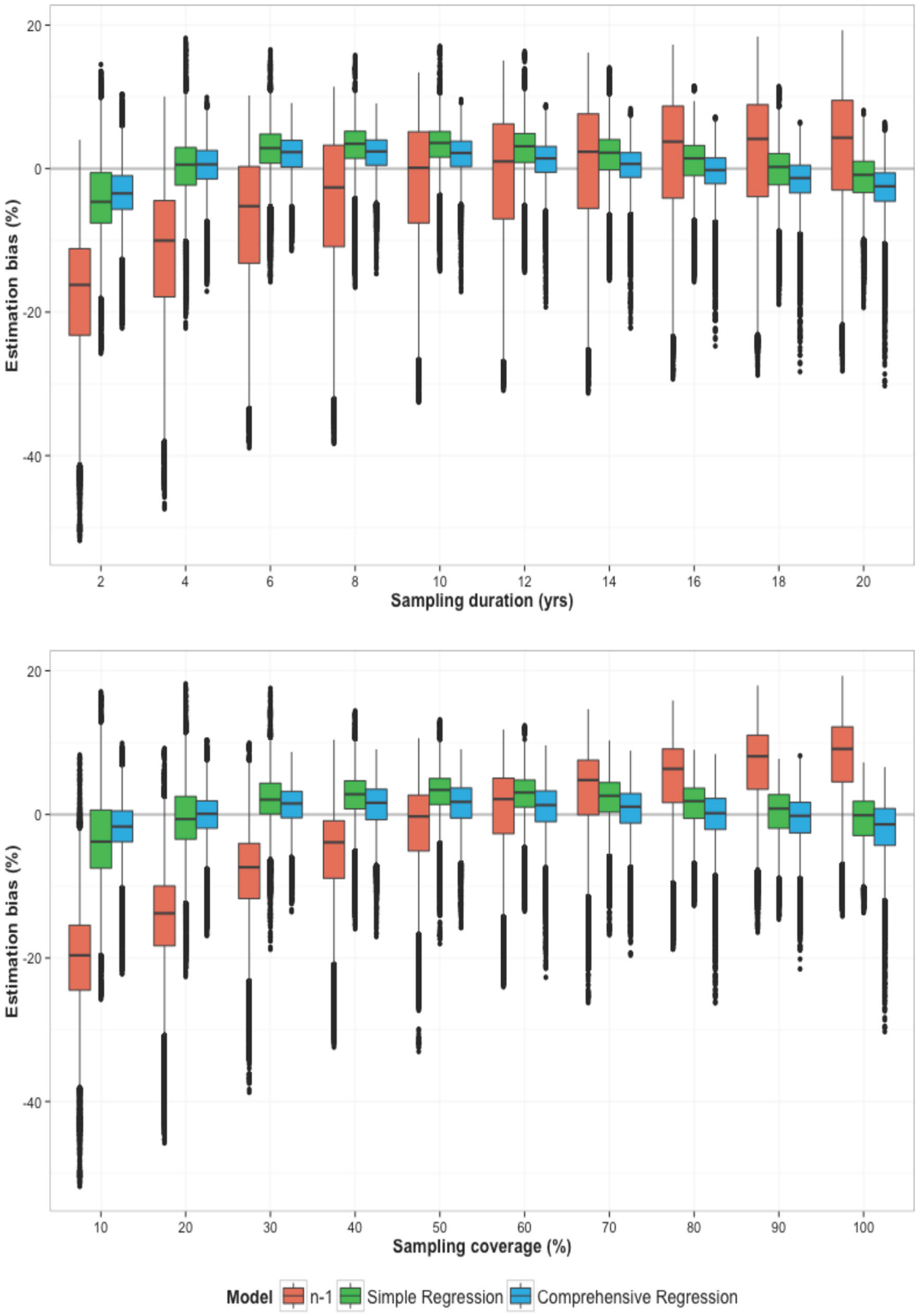
20
0
-20
-40
Estimation bias (%)
2
4
6
8
10
12
14
16
18
20
Sampling duration (yrs)
20
0
-20
-40
Estimation bias (%)
10
20
30
40
50
60
70
80
90
100
Sampling coverage (%)
Model
n-1
Simple Regression
Comprehensive Regression

Fig 4. Estimation bias at different levels of study duration (top) and coverage (bottom), comparing the 'n-1' method to novel regression-based tools. The y-axis presents the (non-absolute) estimation bias [(estimated value – true value) × 100] in the proportion of incident active TB due to recent transmission ("recent transmission proportion"). Estimates from the 'n-1' method are shown in red, and those from the simple and comprehensive regression tools are shown in green and blue, respectively. Boxes show the interquartile range of values from all simulations, and "whiskers" show the 95% confidence intervals, such that narrower boxes correspond to more precise (reproducible) estimates. The '*n-1*' model tends to underestimate the recent transmission proportion at low levels of sample coverage (<50%) and study duration (<10 years), and begins to overestimate the recent transmission proportion as coverage and duration are increased. The regression models are fairly robust to variation of study characteristics, providing more accurate and precise estimates of recent transmission proportion, especially in settings of incomplete coverage and short study duration.

doi:10.1371/journal.pone.0144137.g004

underestimates of the recent transmission proportion, as cases in the same transmission chain were often miscategorised as non-clustered. This pattern is evident in the left side of Fig 4, for sampling duration of less than 10 years (upper panel) or sampling coverage of less than 60% (lower panel), where the '*n*-1' approach (in red bars) increasingly underestimates the recent transmission proportion as data become more incomplete.

Sensitivity Analysis

In analyses of partial rank correlation coefficients, bias in the '*n*-1' estimates was strongly and positively correlated with the duration (PRCC = 0.79) and coverage of fingerprint data (PRCC = 0.89). The regression models, on the other hand, provided more accurate and precise estimates of the recent transmission proportion in the setting of incomplete data collection. For example, among simulations with 30–40% population coverage of molecular data over a 2-to-4 year duration of data collection, the simple regression model had a median estimation bias of 0.2% [-4%, 4%], compared to a substantial underestimation of -18% [-21%, -15%] with the '*n-1*' method. At 70–80% coverage for 10–12 years, by contrast, the median bias in the simple regression estimate was 4% [2.4%, 5.5%], not noticeably different from that of the *'n-1'* method (5% [4%, 7%]). These findings were consistent with the results of a large-scale whole-genome sequencing study in the Karonga district of Malawi [48], in which a sample of TB cases over 15 years with an approximate coverage of 50% resulted an estimate of 38% for the recent transmission proportion, whereas the simple regression estimate was about 32% (further details in Section C.4 of the S1 Appendices). PRCCs for correlation between the simple model output and study duration/coverage were -0.03 and 0.17, respectively (Section C.3 and C.4 of the S1 Appendices).

Variation in the underlying fingerprint replacement rate for individuals with latent TB infection influenced the performance of models, but it did not change the relative performance of the regression-based models versus the '*n*-1' method. At higher fingerprint replacement rates (higher strain heterogeneity), the models tended to underestimate the recent transmission proportion [30,43], but the effect was similar both models (*'n-1'* and regression), such that the regression models provided more precise estimates of recent transmission in all scenarios. Similarly, the models tended to overestimate the recent transmission proportion in settings with low fingerprint heterogeneity (low replacement rate) [49], but this effect was similar for both regression and *'n-1'* techniques. Section D.1 of the S1 Appendices provides more detail. When we studied the performance of models in the remaining 419 simulations with an incidence higher than 450 per 100,000/year, the regression tools continued to outperform the '*n*-1' method (Section D.2 of the S1 Appendices).

Illustrative Case

In the original study (recreated by the calibrated simulation model as described in Section E of the S1 Appendices), the '*n*-1' method provided a close estimate of the underlying recent transmission proportion (<1% bias, asterisk in Fig 5), due to reasonable completeness of data

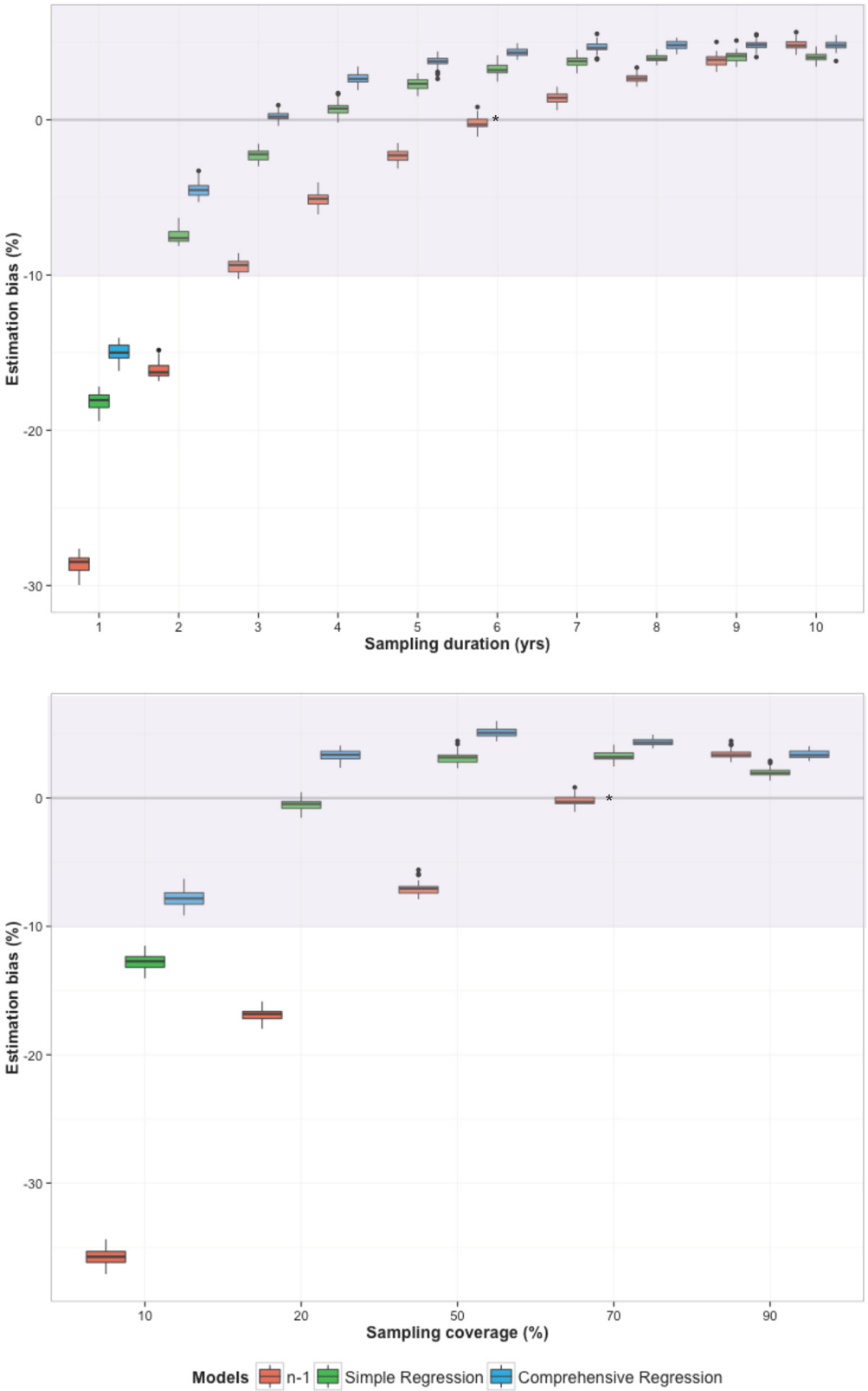
Estimation bias (%)
0
-10
-20
-30
1
2
3
4
5
6
7
8
9
10
Sampling duration (yrs)
Estimation bias (%)
0
-10
-20
-30
10
20
50
70
90
Sampling coverage (%)
Models
n-1
Simple Regression
Comprehensive Regression

Fig 5. Estimation bias in the TB recent transmission proportion in an illustrative high burden case study, comparing the 'n-1' method to the regression-based models at different levels of sampling coverage. The top panel compares the estimation bias resulting from each model at a fixed sampling coverage of 73% (as estimated assuming a 90% case detection proportion, referred to as "Scenario 1" in the text), while varying the duration of data collection. The bottom panel present the results at a fixed study duration of 6 years (as reported in the original study), while varying the coverage of molecular fingerprints at the population level. The asterisk denotes the baseline scenario as reported in reference [23], at which all three methods accurately estimate the recent transmission proportion to within 5%. However, as study duration and population coverage decline, the performance of the 'n-1' method falls dramatically. At either a two-year study duration or a 20% population coverage of molecular data, the '*n-1*' method underestimates the recent transmission proportion by 17% (second column of each figure), whereas both regression tools continue to estimate the recent transmission proportion with a bias of 7% or less. Boxes show the interquartile range of values from all simulations, and "whiskers" show the 95% confidence intervals, such that narrower boxes correspond to more precise (reproducible) estimates. Note that the three bars are jittered at each level of coverage/duration for clarity, but all three methods are performed under the same conditions.

doi:10.1371/journal.pone.0144137.g005

collection (e.g., sample duration of 6 years and coverage of 73% in scenario 1). This performance, however, deteriorated under the assumption that fingerprinting data might be more limited in corresponding programmatic settings. For example, if data were available for six years (as in the original study), but samples were collected only from 30% of diagnosed cases (instead of 73% in the original study), the median bias using the '*n*-1' method inflated to -17% (Fig 5, lower panel, second column, red bars). Similarly, if the study duration was limited to two years, while maintaining the sampling coverage at 73% (per original study), the median bias using the '*n*-1' method was again -17% (Fig 5, upper panel, second column, red bars). In both of these cases, use of the regression tools rather than the *'n-1'* method would have reduced this bias into the range of -7% to +3% (Fig 5, green and blue bars). Similar results were observed in scenario 2 (Fig N of the S1 Appendices), with the regression tools having similar performance to the *'n-1'* method in settings of comprehensive data collection, but showing substantially less bias in settings where data collection was incomplete.

Discussion

The recent transmission proportion is an important indicator of the degree to which observed TB incidence reflects ongoing disease transmission in a given population, with key policy implications including selection of appropriate interventions for disease control. We present here a set of two regression-based tools, the simpler of which is accessible via an easy-to-use web interface (http://modeltb.org/recenttrans/), and each of which removes 60–70% of the bias in estimating the recent transmission proportion in programmatic settings where molecular TB data may be incomplete or short-term in nature. Use of these tools may substantially advance our ability to link programmatic, often sparse molecular data on TB strain clustering to more accurate estimates of recent transmission and thus to more appropriate public health decision-making, without the need to conduct long and extensive molecular epidemiological studies where resources are limited. This meta-modeling technique using a combination of individual-based simulation and regression may also serve as a paradigm with broader application to using population-based (but sparse) molecular data to better understand infectious disease dynamics on longer time scales.

Previous studies have demonstrated the degree of bias in '*n*-1' estimates [32], for example showing that clustering levels will underestimate recent transmission in settings of low data coverage. Our simulations further confirm these results, showing that gaps in sampling cause compensatory underestimates in the recent transmission proportion when using the '*n*-1' method. Our model extends these earlier findings by providing an accessible tool that enables public health decision-makers to input estimates of study duration, sampling coverage, and TB incidence to obtain a more accurate estimate of the recent transmission proportion under local conditions. This new method is particularly relevant in settings of incomplete coverage of fingerprinting data, and accommodates the application of limited data to programmatic decision-

making. Thus, a public health system in a high-burden country with no existing molecular repository may now be able to obtain reasonably accurate estimates of the recent transmission proportion by fingerprinting only a few years of data (either retrospective analysis of existing isolates or prospective collection of isolates over two to three years), and without the need for full population coverage (as long as the sample is representative of an underlying population). As an example, we studied the application of these regression tools in an epidemiological setting discussed in the literature [23] and evaluated the results assuming different levels of access to molecular fingerprinting data. Specifically, we found that, while both the regression tools and '*n-1*' method performed well in the setting of complete data availability, the regression tools markedly reduced bias in estimating the recent transmission proportion when data were incomplete (due either to short duration or low sampling coverage). These tools may therefore enable policy decisions that are more data-driven in settings of high TB burden, even when molecular fingerprint data are relatively sparse.

As with any other modeling study, our analysis has certain limitations. First, to retain strain diversity over time, we implemented an artificial "fingerprint replacement" procedure to capture uncertain patterns of migration, mutation rates, and exogenous infection over time. Our simulations may therefore underestimate the true recent transmission proportion in settings where mutation, migration, and mixing with external populations occur at a high rate, and overestimate the recent transmission proportion in isolated populations where such replacement rarely occurs. Importantly, we varied this "replacement" rate over a wide range in sensitivity analysis and did not see any substantial impact on our results. Second, because of difficulties in setting up appropriately representative closed-population simulations in low-burden settings where the majority of incident TB is imported, we restricted our model to moderate-to-high burden settings. These results should therefore be generalized to lower-burden settings only with caution, and future simulation efforts may be useful in developing a tool that is more appropriate for settings in which immigration drives the majority of incident TB. Third, as our aim was to provide a generalizable, transparent, user-friendly platform, we excluded complexities such as age structure, HIV coinfection, and heterogeneous mixing. Prior studies have suggested that these heterogeneities may affect the estimates of clustering (and therefore recent transmission proportion) in different ways (for example, underestimating clustering among younger individuals and overestimating clustering in older individuals) [33,43,50]. Future efforts could incorporate these assumptions into more complex models to evaluate the residual amount of bias introduced by such simplifications (which should apply equally to the regression and *'n-1'* approaches). Finally, since our goal here was to generate a method that more accurately represents the truth (especially as the resolution of genotyping data is anticipated to improve over time), we assumed a genotyping method with perfect resolution between strains. To the extent that less-discriminatory methods are used for fingerprinting, estimates of clustering from all methods are likely to be positively biased, resulting in overestimates of the recent transmission proportion.

In summary, we have created novel tools, using regression and an individual-based simulation model, to better estimate the proportion of incident TB due to recent transmission in high-burden settings. These tools remove 60–70% of the bias intrinsic to the most commonly used method at present (the '*n*-1' method), and the simple tool is easily accessible to epidemiologists and public health officials via a web-based user interface. Such approaches may have broader applicability to the estimation of clustered transmission of other infectious diseases as well. As we seek to accelerate progress in the fight against these diseases worldwide, better estimates of the recent transmission proportion in subpopulations will be critical to developing evidence-based public-health approaches that appropriately target those hotspots of recent transmission. For example, more accurate estimates of the recent progression proportion could

assist local-level officials in deciding whether to allocate resources to interventions (e.g., contact investigation, improved diagnosis, and case finding) more appropriate for epidemics driven by recent transmission, or to those (e.g., preventive therapy) more targeted toward preventing reactivation. Tools such as these can help guide decision-makers to develop more effective policies and interventions in settings where data are often incomplete and long studies are impractical.

Supporting Information

S1 Appendices.
(DOCX)

Author Contributions

Conceived and designed the experiments: PK DWD. Performed the experiments: PK. Analyzed the data: PK DWD. Contributed reagents/materials/analysis tools: PK DWD JP. Wrote the paper: PK DWD. Revised the manuscript and contributed intellectual content: PK BM WDK ASA DWD.

References

1. Lin PL, Flynn JL. Understanding latent tuberculosis: a moving target. J Immunol. American Association of Immunologists; 2010; 185: 15–22. doi: 10.4049/jimmunol.0903856 PMID: 20562268
2. Parrish NM, Dick JD, Bishai WR. Mechanisms of latency in Mycobacterium tuberculosis. Trends Microbiol. Elsevier; 1998; 6: 107–112. doi: 10.1016/S0966-842X(98)01216-5 PMID: 9582936
3. Jasmer RM, Nahid P, Hopewel philip C. Latent tuberculosis infection. N Engl J Med. 2002; 347: 1860–1866. PMID: 12466511
4. Alland D, Kalkut GE, Moss AR, McAdam RA, Hahn JA, Bosworth W, et al. Transmission of tuberculosis in New York City—an analysis by DNA fingerprinting and conventional epidemiologic methods. N Engl J Med. 1994; 330: 1710–1716. PMID: 7993412
5. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, et al. The Epidemiology of Tuberculosis in San Francisco—A Population-Based Study Using Conventional and Molecular Methods. N Engl J Med. 1994; 330: 1703–1709. PMID: 7910661
6. Geng E, Kreiswirth B, Driver C, Li J, Burzynski J, DellaLatta P, et al. Changes in the transmission of tuberculosis in New York City from 1990 to 1999. N Engl J Med. 2002; 346: 1453–1458. PMID: 12000815
7. Morrison J, Pai M, Hopewell PC. Tuberculosis and latent tuberculosis infection in close contacts of people with pulmonary tuberculosis in low-income and middle-income countries: a systematic review and meta-analysis. Lancet Infect Dis. Elsevier; 2008; 8: 359–368. doi: 10.1016/S1473-3099(08)70071-9 PMID: 18450516
8. Badri M, Wilson D, Wood R. Effect of highly active antiretroviral therapy on incidence of tuberculosis in South Africa: a cohort study. Lancet. 2002; 359: 2059–64. doi: 10.1016/S0140-6736(02)08904-3 PMID: 12086758
9. Kasaie P, Mathema B, Azman A, Kelton D, Dowdy D. Better Than "n-1" Model: Estimating the proportion of tuberculosis recent transmission via simulation. 2014.
10. Azman AS, Golub JE, Dowdy DW. How much is tuberculosis screening worth? Estimating the value of active case finding for tuberculosis in South Africa, China, and India. BMC Med. 2014; 12: 216. doi: 10.1186/s12916-014-0216-0 PMID: 25358459
11. Mathema B, Lewis JJ, Connors J, Chihota VN, Shashkina E, Meulen M van der, et al. Molecular epidemiology of Mycobacterium tuberculosis among South African gold-miners. Ann Am Thorac Soc. 2014;
12. Horsburgh CR Jr. Priorities for the treatment of latent tuberculosis infection in the United States. N Engl J Med. Mass Medical Soc; 2004; 350: 2060–2067. PMID: 15141044
13. Golub JE, Astemborski J, Ahmed M, Cronin W, Mehta SH, Kirk GD, et al. Long-term effectiveness of diagnosing and treating latent tuberculosis infection in a cohort of HIV-infected and at-risk injection drug users. J Acquir Immune Defic Syndr. 2008; 49: 532–7. doi: 10.1097/QAI.0b013e31818d5c1c PMID: 18989223

14. McKenna MT, McCray E, Onorato I. The epidemiology of tuberculosis among foreign-born persons in the United States, 1986 to 1993. N Engl J Med. 1995; 332: 1071–1076. PMID: 7898526

15. Genewein A, Telenti A, Bernasconi C, Schopfer K, Bodmer T, Mordasini C, et al. Molecular approach to identifying route of transmission of tuberculosis in the community. Lancet. 1993; 342: 841–844. PMID: 8104275

16. Mathema B, Kurepina NE, Bifani PJ, Kreiswirth BN. Molecular epidemiology of tuberculosis: current insights. Clin Microbiol Rev. 2006; 19: 658–85. doi: 10.1128/CMR.00061-05 PMID: 17041139

17. Houben RMGJ, Yates TA, Moore DAJ, McHugh TD, Lipman M, Vynnycky E. Estimated Rate of Reactivation of Latent Tuberculosis Infection in the United States, Overall and by Population Subgroup. Am J Epidemiol. 2014;kwu187.

18. Meissner JS, Crossa A, R. Chaisson JG, Ahuja SD. Contribution Of Country Of Birth To The Transmission Of Mycobacterium Tuberculosis In New York City (nyc). Am J Respir Crit Care Med. 2012; 185: Meissner, J. Sullivan, A. Crossa, R. Chaisson, J.

19. Driver CR, Kreiswirth B, Macaraig M, Clark C, Munsiff SS, Driscoll J, et al. Molecular epidemiology of tuberculosis after declining incidence, New York City, 2001–2003. Epidemiol Infect. 2007; 135: 634–643. PMID: 17064454

20. Anderson LF, Tamne S, Brown T, Watson JP, Mullarkey C, Zenner D, et al. Transmission of multidrug-resistant tuberculosis in the UK: a cross-sectional molecular and epidemiological study of clustering and contact tracing.". Lancet Infect Dis. 2014; 14: 406–415. doi: 10.1016/S1473-3099(14)70022-2 PMID: 24602842

21. van Deutekom H, Hoijng SP, de Haas PEW, Langendam MW, Horsman A, van Soolingen D, et al. Clustered tuberculosis cases: do they represent recent transmission and can they be detected earlier? Am J Respir Crit Care Med. American Thoracic Society; 2004; 169: 806–10. doi: 10.1164/rccm.200306-856OC PMID: 14684559

22. van Soolingen D, Borgdorff MW, de Haas PE, Sebek MM, Veen J, Dessens M, et al. Molecular epidemiology of tuberculosis in the Netherlands: a nationwide study from 1993 through 1997. J Infect Dis. 1999; 180: 726–36. doi: 10.1086/314930 PMID: 10438361

23. Verver S, Warren RM, Munch Z, Richardson M, van der Spuy GD, Borgdorff MW, et al. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. Lancet. 2004; 363: 212–4. doi: 10.1016/S0140-6736(03)15332-9 PMID: 14738796

24. Middelkoop K, Mathema B, Myer L, Shashkina E, Whitelaw A, Kaplan G, et al. Transmission of Tuberculosis in a South African Community With a High Prevalence of HIV Infection. J Infect Dis. 2014; JIU403.

25. Bishai WR, Graham NM, Harrington S, Pope DS, Hooper N, Astemborski, Jacqueline Sheely L, et al. Molecular and geographic patterns of tuberculosis transmission after 15 years of directly observed therapy. Jama. 1998; 280: 1679–1684. PMID: 9831999

26. Schürch AC, van Soolingen D. DNA fingerprinting of Mycobacterium tuberculosis: from phage typing to whole-genome sequencing. Infect Genet Evol. 2012; 12: 602–9. doi: 10.1016/j.meegid.2011.08.032 PMID: 22067515

27. Roetzer A, Diel R, Kohl TA, Rückert C, Nübel U, Blom J, et al. Whole genome sequencing versus traditional genotyping for investigation of a Mycobacterium tuberculosis outbreak: a longitudinal molecular epidemiological study. Neyrolles O, editor. PLoS Med. Public Library of Science; 2013; 10: e1001387. doi: 10.1371/journal.pmed.1001387 PMID: 23424287

28. Houben RM, Glynn JR. A systematic review and meta-analysis of molecular epidemiological studies of tuberculosis: development of a new tool to aid interpretation. Trop Med Int Heal. 2009; 14: 892–909.

29. Glynn JR, Bauer J, De Boer AS, Borgdorff MW, Fine PEM, Godfrey-Faussett P, et al. Interpreting DNA fingerprint clusters of Mycobacterium tuberculosis Position Paper. Int J Tuberc Lung Dis. 1999; 3: 1055–1060. PMID: 10599007

30. Glynn JR, Vyonycky E, Fine PEM. Influence of sampling on estimates of clustering and recent transmission of Mycobacterium tuberculosis derived from DNA fingerprinting techniques. Am J Epidemiol. 1999; 149: 366–371. PMID: 10025480

31. Murray M, Alland D. Methodological problems in the molecular epidemiology of tuberculosis. Am J Epidemiol. 2002; 155: 565–571. PMID: 11882530

32. Murray M. Sampling bias in the molecular epidemiology of tuberculosis. Emerg Infect Dis. 2002; 8: 363. PMID: 11971768

33. Vynnycky E, Nagelkerke N, Borgdorff MW, Van Soolingen D, Van Embden JDA, Fine PEM, et al. 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. Epidemiol Infect. Cambridge Univ Press; 2001; 126: 43–62. PMID: 11293682

34. Grabowski MK, Lessler J, Redd AD, Kagaayi J, Laeyendecker O, Ndyanabo A, et al. The role of viral introductions in sustaining community-based HIV epidemics in rural Uganda: evidence from spatial clustering, phylogenetics, and egocentric transmission models. PLoS Med. 2014; 11: e1001610. doi: 10.1371/journal.pmed.1001610 PMID: 24595023

35. Mahony AA, Donnan EJ, Lester RA, Doyle JS, Knox J, Tracy SL, et al. Beyond injecting drug use: investigation of a Victorian cluster of hepatitis C among HIV-infected men who have sex with men.". Med J Aust. 2013; 198: 210–214. PMID: 23451966

36. Kouyos RD, Rauch A, Böni J, Yerly S, Shah C, Aubert V, et al. Clustering of HCV coinfections on HIV phylogeny indicates domestic and sexual transmission of HCV.". Int J Epidemiol. 2014; dyt276.

37. Kasaie P, Kelton WD, Dowdy DW. Estimating proportion of tuberculosis recent transmission via simulation. In: Tolk A, Diallo SD, Ryzhov IO, Yilmaz L, Buckley S, Miller JA, editors. Proceedings of Winter Simulation Conference,IEEE. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.; 2014. pp. 1469–1480.

38. Vynnycky E, Fine PE. The natural history of tuberculosis: the implications of age-dependent risks of disease and the role of reinfection. Epidemiol Infect. 1997; 119: 183–201. PMID: 9363017

39. Tiemersma EW, van der Werf MJ, Borgdorff MW, Williams BG, Nagelkerke NJD. Natural history of tuberculosis: duration and fatality of untreated pulmonary tuberculosis in HIV negative patients: a systematic review. PLoS One. Public Library of Science; 2011; 6: e17601.

40. World Health Organization. Global TB Report [Internet]. 2013. Available: http://www.who.int/tb/publications/global_report/en/

41. Sutherland I, Švandová E, Radhakrishna S. The development of clinical tuberculosis following infection with tubercle bacilli: 1. A theoretical model for the development of clinical tuberculosis following infection, linking from data on the risk of tuberculous infection and the incidence of clinic. Tubercle. Elsevier; 1982; 63: 255–268. PMID: 6763793

42. Andrews JR, Noubary F, Walensky RP, Cerda R, Losina E, Horsburgh CR. Risk of progression to active tuberculosis following reinfection with Mycobacterium tuberculosis. Clin Infect Dis. 2012; 54: 784–91. doi: 10.1093/cid/cir951 PMID: 22267721

43. Murray M. Determinants of cluster distribution in the molecular epidemiology of tuberculosis. Proc Natl Acad Sci. 2002; 99: 1538–1543. PMID: 11818527

44. Fok A. YN, Schulzer M, FitzGerald MJ. Risk factors for clustering of tuberculosis cases: a systematic review of population-based molecular epidemiology studies [Review Article]. Int J Tuberc Lung Dis. 2008; 12: 480–492. PMID: 18419882

45. 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: 458–465. doi: 10.3201/eid1803.111107 PMID: 22377473

46. World Health Organization. WHO TB burden estimates [Internet]. 2013. Available: http://www.who.int/tb/country/data/download/en/

47. Sanchez MA, Blower SM. Uncertainty and sensitivity analysis of the basic reproductive rate: tuberculosis as an example. Am J Epidemiol. 1997; 145: 1127–1137. PMID: 9199543

48. Guerra-Assunção JA, Crampin AC, Houben RMGJ, Mzembe T, Mallard K, Coll F. Large-scale whole genome sequencing of M. tuberculosis provides insights into transmission in a high prevalence area. Elife. 2015; 4: e05166.

49. Benedetti A, Menzies D, Behr MA, Schwartzman K, Jin Y. How close is close enough? Exploring matching criteria in the estimation of recent transmission of tuberculosis. Am J Epidemiol. 2010; 172: 318–26. doi: 10.1093/aje/kwq124 PMID: 20576754

50. Hollm-Delgado M-G. Molecular epidemiology of tuberculosis transmission: Contextualizing the evidence through social network theory. Soc Sci Med. 2009; 69: 747–753. doi: 10.1016/j.socscimed.2009.06.043 PMID: 19625118