WEB MATERIAL

Modeling Missing Cases and Transmission Links in Networks of Extensively Drug-Resistant Tuberculosis in KwaZulu-Natal, South Africa

Kristin N. Nelson, Neel R. Gandhi, Barun Mathema, Benjamin A. Lopman, James C. M. Brust, Sara C. Auld, Nazir Ismail, Shaheed Vally Omar, Tyler S. Brown, Salim Allana, Angie Campbell, Pravi Moodley, Koleka Mlisana, N. Sarita Shah, and Samuel M. Jenness

Table of Contents

Web Appendix 1: Empirical Data

The Transmission Study of XDR TB (TRAX Study) is a cross-sectional study that enrolled 404 XDR TB cases from KwaZulu-Natal province, South Africa from 2011–2014. This study collected clinical, demographic and social network data from enrolled cases. The primary aim of this study was to estimate the proportion of XDR TB cases due to transmission, as compared to those due to acquired resistance. The major finding of this study was that at least 70% of XDR cases in this settings are due to transmission.[1]

Briefly, the diagnostic XDR Mtb isolate was obtained for all participants and re-cultured on Löwenstein-Jensen slants. We conducted population sweeps, extracted genomic DNA, and prepared sequencing libraries using Nextera DNA kits (Illumina, San Diego, CA). Raw paired-end sequencing reads were generated on the Illumina (MiSeq) platform and aligned to the H37Rv reference genome (NC_000962.3) using the Burrows-Wheeler Aligner. All isolates had reads covering >99% of the reference genome, and the lowest mean coverage depth for any isolate was 15X. SNPs were detected using standard pairwise resequencing techniques (Samtools v0.1.19) against the reference and filtered for quality, read consensus (>75% reads for the alternate allele) and proximity to indels (>50 base-pairs from any indel). SNPs in or within 50 base pairs of hypervariable PPE/PE gene families, repeat regions, and mobile elements were excluded.[2] The empirical transmission network was created from 344 cases with available whole genome sequencing results of their *Mtb* isolates. Sequencing data are available on the NCBI Sequence Read Archive (BioProject: PRJNA476470).

We created sequencing-based networks using pairwise differences between *Mtb* sequences. We considered fewer than 5 single nucleotide polymorphisms a transmission link and constructed an undirected network in which each node represented a case and each edge represented a transmission link. We considered several SNP thresholds, as the appropriate threshold to define a direct transmission event between two cases may vary by setting, study design, and SNP calling pipeline used (see Sensitivity Analyses section below).[3] Notably, our empirical network also makes the assumption that individuals are infected with only the sequenced TB variant. While multiple infections are common in high TB incidence settings, this assumption is likely reasonable given the relatively low force of infection of XDR TB.

In this empirical undirected network, the maximum degree was 62, there were 162 (47%) of cases with no links (degree = 0), and 62 (18%) of cases with 10 or more links (degree ≥ 10). See Web Tables 1 and 3 for more descriptive characteristics of the empirical network defined by 5 SNPs and 3 SNPs, respectively.

To define the likelihood of being linked in the modeled networks based on a particular attribute, we defined the expected mean degree of cases with a given attribute as compared to a reference group. (For example, we defined the expected mean degree of HIV-positive relative to HIV-negative cases based on estimates from the literature on the relative infectiousness of HIV-positive as compared to HIV-negative cases.)

To calculate the target statistics for each nodefactor term in the network model, we multiplied the relative mean degree for each attribute by the overall mean degree of the network. To scale this according to the number of nodes in the network, we multiplied that value by the proportion of nodes in that network with the attribute.

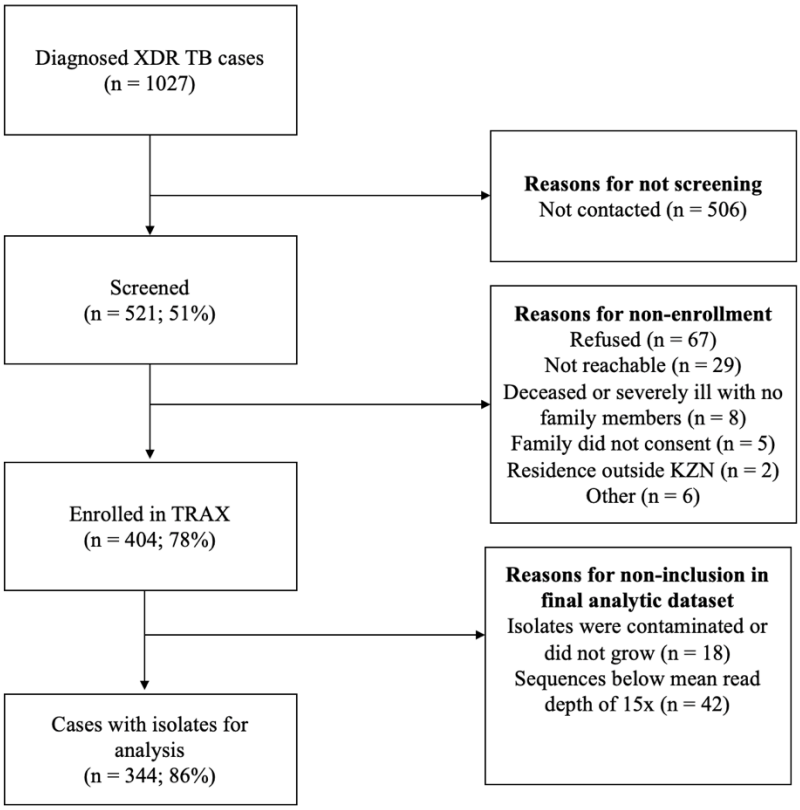
As a brief example: Assume we want to model a network with 100 TB cases that has an overall mean degree of 5. Assume this network includes 75 HIV-positive and 25 HIV-negative TB cases.

Let's assume the mean degree of HIV-positive cases is 2 and for HIV-negative cases is 1, giving a relative mean degree for HIV-positive cases of 2/3. We multiply: 5 × 100 to get the total number of edges in this network (n = 500 for a directed network, n = 250 for an undirected network). Then, we know that HIV-positive cases are twice as likely to have an edge as HIV-negative cases, so 250 × 2/3 gives 167 total number of edges associated with HIV-positive cases in the network.

Models with target statistics specified for all levels of every attribute did not easily converge, so we reduced the number of target statistics for attributes with more than four categories. For these variables, we used the 3–4 categories corresponding to the highest number of edges as target statistics to parameterize models. Using these target statistics, we simulated full transmission networks from each model.

In order to simulate using these target statistics, we had to specify a mean degree of the full transmission network. In theory, this mean degree would be the mean number of secondary transmissions per case, plus 1 (the link to their source case). The effective reproduction number of TB has been estimated to range from 1 to 5, and a recent study on superspreading in TB estimated the mean of the secondary case distribution to be.[4-6] However, these studies have been done primarily in high-income countries with low TB incidence, and in South Africa, diagnostic delays may lead to long infectious periods and a higher rate of transmission. So, we simulated full transmission networks assuming a range of mean degrees from 1 to 10. In addition, we also examined networks with mean degrees of 15 and 20 to better understand how the network behaved under higher mean degrees.

Web Figure 1. Transmission of XDR TB study in KwaZulu-Natal, South Africa, 2011–2014, enrollment and study inclusion



Web Appendix 2: Model details

This technical appendix describes the models used in the associated manuscript, including their conceptual basis and parameterization as well as simulation procedures and statistical analysis.

1.1 Model framework

The network models in this study were used to represent and simulate transmission networks of active tuberculosis (TB) cases. Links, or edges, in modeled networks represent a transmission event that occurred between two cases in the network. We considered the structure of the complete network as a joint function of the empirical data and assumptions about missingness.

Modeled networks do not involve individuals (1) *infected* with TB but whom did not progress to active disease or (2) exposed contacts of TB cases. Rather, modeled networks reflect all transmission events observed from a sample of cases enrolled in our transmission study of XDR TB, which enrolled patients over a four-year period. Note that this network does not include transmission events that occurred either before the study period began or after it ended.

Moreover, this network is comprised of cases with extensively drug-resistant (XDR) TB. Importantly, not all cases of XDR TB were infected with XDR TB; rather, they acquired resistance as a result of inadequate treatment of a more drug-susceptible strain of TB. For simplicity, we ignore this feature of drug-resistant TB epidemiology when modeling the transmission networks in this study. Moreover, it has been shown that the vast majority of XDR TB are due to transmission of already-resistant strains, rather than acquired resistance. Of note, the full networks that we modeled *do* include unconnected nodes, or cases, which could represent individuals who acquired their XDR through inadequate treatment and therefore would not be connected to a source case in the network. So, our models indirectly account for the possibility that some XDR TB is acquired rather than transmitted.

In the model, each case in the network was assigned specific clinical and demographic attributes according to pre-defined distributions. Each attribute is represented by a 'nodefactor' term in the network model. (Some 'nodefactor' terms represented the joint distributions of two attributes, see Joint Distributions section.) This allowed the number of links to vary by an individual's attributes. We defined 'target statistics' for the number of links, or edges, attributed to cases with a given attribute in the network. (For example, the target statistic for the HIV nodefactor term was the number of edges involving HIV-positive nodes, or the number of transmission events in the network involving HIV-positive cases.)

For attributes with well-studied effects on infectiousness (i.e., smear status), we used estimates from the literature to define target statistics for the corresponding nodefactor term in our models. In the absence of available data from the literature, we used data collected from our transmission study. (Ultimately, we used data from our transmission study for only two parameters: the relationship between transmission risk and cough duration and the distribution of XDR TB strain type in KwaZulu-Natal. This information was not readily available outside our study.) The attributes assigned to each case collectively influence the number of other cases to whom that case is connected in the network. (See the Empirical Data section for more detail.)

1.2 Model software

All models were programmed in R. The code used to create and analyze these models is available on GitHub (https://github.com/kbratnelson/tb-ergms).

We used the *ergm* R package, which requires the *statnet* suite of software.

Web Appendix 3: Clinical measures

3.1 Cough duration

We categorized cough duration by month. The distribution of cough duration and target statistics for the mean degree in each group are below:

Web Table 1.

Cough Duration	No. (%)	Mean Degree (Source: TRAX)
No cough*	128 (37)	5.0
1 month*	60 (17)	6.5
2 months*	51 (15)	8.1
3 months*	72 (21)	8.1
4 months	16 (5)	4.6
5 months	17 (5)	3.7

* Target statistics defined for these categories in ERGMs.

As there was little information on the effect of cough duration on transmission in the literature, we used the mean degree from the empirical TRAX network to define target statistics for modeled networks. We used target statistics for the largest categories 'No cough', '1 month', '2 months', and '3 months' to fit models.

3.2 Smear status

Although information on both smear status and smear grade were available in TRAX, we chose to use only smear status (smear-positive and smear-negative) to reduce the number of parameters in the model. The marginal distribution of smear status that we used to parameterize models is below:

Web Table 2.

Smear Status	No. (%)	Mean Degree (Source: Abu-Raddad et al.[11])
Negative	109 (32)	1
Positive	235 (68)	4

These parameters are based on relative infectiousness estimates by Abu-Raddad et al. that we normalized, assuming a mean degree of 1 in the smear-negative group. We used the joint distribution of age and smear status for model target statistics; see Joint Distributions section.

3.3 HIV

Although both HIV status and information on virologic suppression were available in TRAX, we chose to use only smear status (smear-positive and smear-negative) to reduce the number of parameters in the model. The marginal distribution of HIV status that we used to parameterize models is shown below.

Web Table 3.

HIV Status	No. (%)	Mean Degree
Negative	78 (23)	1
Positive	266 (77)	1

Since there is conflicting evidence as to whether HIV-positive or HIV-negative individuals are more infectious, we chose to assume no difference in infectiousness and thus the mean degree for HIV-positive and HIV-negative individuals was assumed to be the same. We used the joint distribution of age and HIV status for model target statistics; see Joint Distributions section.

3.4 Mtb *strain type*

The dominant strain of XDR TB in KwaZulu-Natal is the LAM4 strain. There is some evidence that this strain may be unique from other in terms of its transmission and evolutionary rate, so we accounted for this in our models.[12,13] We categorized *Mtb* strains into LAM4 or non-LAM4.

Web Table 4.

Mtb Strain	No. (%)	Mean Degree (Source: TRAX)
LAM4*	259 (75)	8.3
Non-LAM4	85 (25)	0.2

* Target statistics defined for these categories in ERGMs.

Since there was little direct evidence in the literature about the relative infectiousness of LAM4 and non-LAM4 strains of XDR TB, we used the relative mean degrees estimated from the empirical TRAX network for modeled networks.

Web Appendix 4: Demographic measures

4.1 Age

We categorized age into four groups: 0–15, 16–34, 35–54, and ≥55 years.

Web Table 5.

Age Category	No. (%)	Mean Degree (Source: Wood et al.[14])
0–15	12 (3)	1
16–34	171 (50)	1.58
35–54	134 (39)	0.98
≥55	27 (8)	0.75

The mean degree parameters are based on relative infectiousness estimates by Wood et al. that we normalized assuming a mean degree of 1 in the 0–15 age group.[14] We used the joint distribution of age/HIV status and age/smear status for model target statistics; see Joint Distributions section.

Web Appendix 5: Joint distributions

5.1 Age and smear status

Web Table 6.

Age Category	Smear Status	No. (%)	Mean Degree
0–15	Negative	7 (2)	1.00
16–34*	Negative	44 (13)	1.58
35–54*	Negative	40 (12)	0.98
≥55	Negative	18 (5)	0.75
0–15	Positive	5 (1)	4.00
16–34*	Positive	107 (31)	6.32
35–54*	Positive	79 (23)	3.92
≥55	Positive	7 (2)	3.00

* Target statistics defined for these categories in ERGMs.

We calculated the mean degree by multiplying the relative infectiousness measures for smear status and age group from the Tables in Section 5.2 and 6.1, respectively. (We assumed independence of the two measure of infectiousness.) We used target statistics for the largest categories, '16-34, smear-negative', '35-54, smear-negative', '16-34, smear-positive', and '35-54, smear-positive' as target statistics for network models.

5.2 Age and HIV

Web Table 7.

Age Category	HIV Status	No. (%)	Mean Degree
0–15	Negative	5 (1)	1.00
16–34	Negative	41 (12)	1.58
35–54	Negative	15 (4)	0.98
≥55	Negative	17 (5)	0.75
0–15	Positive	7 (2)	1.00
*16–34	Positive	130 (38)	1.58
*35–54	Positive	119 (35)	0.98
≥55	Positive	10 (3)	0.75

* Target statistics defined for these categories in ERGMs.

We calculated the mean degree by multiplying the relative infectiousness measures for HIV status and age group from the Tables in Section 5.3 and 6.1, respectively. (We assumed independence of the two measure of infectiousness.) We used target statistics for the largest categories, '16-34, HIV-positive', and '35-54, HIV-positive' as target statistics for network models.

5.3 Other (Smear status and HIV)

Although smear-negative disease tends to be more common among HIV-positive TB cases, we did not find this association in the empirical data. The proportion of cases with HIV was nearly equivalent among smear-positive and smear-negative cases and the proportion of smear-positive cases was nearly equivalent among HIV-positive and HIV-negative cases. Thus, we chose not to represent the joint distribution of smear status and HIV in model target statistics.

Web Appendix 6: Sensitivity analyses

6.1 Genomic threshold for transmission

Since the threshold for defining genomic evidence of transmission is not well-defined, we also defined an empirical network using a more stringent threshold of 3 pairwise SNP differences. This resulted in no changes to modeled networks but did change the target statistics we attempted to 'match' with modeled, sampled networks. The differences in the empirical networks defined by different SNP thresholds can be examined by comparing Table 1 in the main manuscript (5-SNP threshold) and Web Table 1 (3-SNP threshold). The target statistics for both networks are shown in Tables 3 and 4.

6.2 Full network size

To simulate full networks, we needed to make assumptions about the true size of the full network, that is, the number of cases involved in XDR TB transmission over the time period 2011–2014. We estimated the number of diagnosed and undiagnosed XDR TB cases in KwaZulu-Natal province using data from the South African National Tuberculosis Drug Resistance Survey.[9] We then used active case-finding studies to estimate the proportion of TB cases in South Africa that are undiagnosed.[10]

332,783 TB cases in South Africa in 2014
Proportion of cases with pulmonary TB (infectious form) = 0.89
Proportion of cases in KwaZulu-Natal Province (area of study) = 0.31
Proportion of cases with XDR = 0.005

332,783 × (0.31) × (0.89) × (0.005) × 4 years = 1,836 cases (736 – 2,572)

Accounting for underdiagnosis of TB cases[10], multiply by a factor of 2:

1,836 cases × 2 = 3,672 cases (1,472 – 5,144)

For our primary analysis, we estimated a total number of XDR TB cases on the lower end of this range (n = 2,000) but explored the impact of changing network size (see Sensitivity Analyses section below).

Given the uncertainty around the total number of XDR TB cases contributing to transmission, we considered several other sizes for the full network. We assumed that the full network may be larger than 2,000 cases (n = 4,000 cases), or that it may be smaller (n = 1,500 cases, n = 500 cases). We compared the results from these networks to our main models, which assumed a full network size of 2,000 cases.

Web Appendix 7: Defining models using missing case assumptions

We defined models and simulated full transmission networks under scenarios which made different assumptions about cases missing from the empirical TRAX network.

7.1 Cases missing at random

We assumed that cases missing at random would result in missing transmission links randomly across the network. We simulated full networks with mean degrees of 2, 5, 8, 10, 15, 20, 50, 100, and 200. Results from models of networks with mean degree of 2 through 20 are presented in the main manuscript and models with mean degree greater than 20 are presented in the supplemental results.

The degree of a given case in the network is the sum of the links to that case, (in theory, one, representing the source case), and all forward transmission links from that case. Since only cases of active disease are represented in the network, the mean degree therefore roughly corresponds to the mean number of secondary cases caused by a TB case in the network (less one, corresponding to the source case of their infection) *who progressed to disease during the study period*.

It is important to note that the modeled and simulated networks are undirected, meaning that the direction of transmission is not indicated. Parameterizing models to include both risk factors for infection and transmission was beyond the scope of this project but would be a useful extension of the current model framework.

7.2 Cases missing by level of connectivity

We assumed two opposing scenarios: that cases who were highly connected in the full network were more likely to be sampled, and that cases that were poorly connected in the full network were more likely to be sampled. To simulate the former scenario, we created and used sampling weights proportional to each cases' degree in the full network; to simulate the latter scenario, we created and used sampling weights inversely proportional to degree in the full network.

We sampled cases using this method in full networks with various mean degrees (2, 5, 8, 10, 15, 20).

7.3 Differential sampling by smear status

We assumed that smear-positive cases were more likely to be sampled. To do this, we modified the distribution of smear status in the full network relative to the empirical TRAX network. In the TRAX Study, 68% of XDR TB cases were smear-positive. To estimate the effect of oversampling smear-positive cases, we assumed that the proportion of smear-positive was smaller in the full network than we observed in TRAX.

We modeled full networks in line with this scenario at various mean degrees (2, 5, 8, 10, 15, 20) and sampled 350 cases from each modeled network.

7.4 Unmeasured factor contributing to transmission

We hypothesized that a factor contributing strongly to transmission risk but that was not accounted for in our model (a 'superspreading' factor) might have a substantial impact on full

and sampled network structure.[7,8] In a recent study examining superspreading behavior among TB cases, the mean number of secondary infections per index case was 0.77, and the 90th percentile of the distribution of secondary infections was 10 infections per index.[4] Consistent with these findings, we hypothesized that a 'superspreading' factor present in a minority of the population might increase transmission by at least 10 times (10x), that is, cases with this factor would be responsible for 10 times as many transmission events as those lacking this factor.

We created a 'nodefactor' term in the model for this unmeasured, superspreading factor. We tested superspreading factors ranging from 10x to 40x. To define the strength of a factor, we increased the degree in the network of cases with this factor relative to cases without the latent factor. The number of links in the overall network remained the same, but adding a latent factor caused them to be distributed differently (more among 'superspreaders', fewer among non-superspreaders.) We varied the strength of the superspreading factor up to 40x to explore the effects on the distribution at higher values for latent factor strength.

In addition, we varied the prevalence (0.10-0.30) of this latent factor. In the same study referenced above, approximately 10% of TB cases were associated with superspreading events. Thus, we varied prevalence starting at 10% to account for the possibility of high-transmitting cases may be undersampled, consistent with our findings from this analysis.

The results of models assuming the strongest effect of the factor (x40) in the smallest proportion (10%) of cases are presented in the main manuscript.

Web Appendix 8: Simulation and sampling methods

From each network model, we simulated 1,000 networks. We specified the following parameters of the Markov Chain Monte Carlo (MCMC) algorithm: we set the number of burn-in simulations as 100,000, the MCMC interval as 5,000, and the MCMC sample size as 10,000.

We ensured that the MCMC algorithm used to estimate parameters for each model converged appropriately by checking for adequate mixing of the MCMC chain and sufficient exploration of parameter space using the mcmc.diagnostics function in the *ergm* package.

We sampled 350 cases from each simulated, full network, mimicking sampling 350 cases in our TRAX Study from the larger population of XDR TB cases. We compared the degree distributions of modeled, sampled networks to that of the empirical TRAX network.

We attempted to 'match' the following quantiles of the empirical degree distribution: (1) 10^{th} percentile; (2) 25^{th} percentile; (3) median (50^{th} percentile), (4) 75^{th} percentile, and (5) maximum (100^{th} percentile).

To statistically compare the degree distributions of the modeled and empirical networks, we used a modified Kolmogorov-Smirnov (K-S) test statistic calculated by bootstrapping techniques using the ks.boot function (in the Matching package) in R. We considered a two-sided alternative hypothesis and used 1,000 bootstraps to calculate *P* values. [15]{Janssen, 1994 #935}

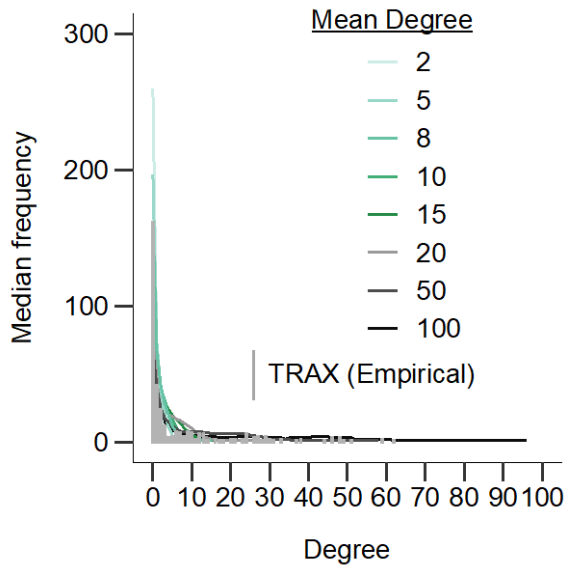
We calculated a *P* value comparing each of 1,000 simulated networks against the empirical network, creating a distribution of K-S *P* values. We report the median *P* value from this distribution in Tables 2 and 3.

Web Table 8. Descriptive characteristics, sequencing-based network of XDR TB cases in the TRAX Study (≤5-SNP threshold), 2011–2014

	No. (%)	Mean
By attribute		
HIV status		
HIV-negative	78 (23)	6.2
HIV-positive, undetectable viral load	133 (39)	6.7
HIV-positive, detectable viral load	133 (39)	5.9
Cough duration		
No cough	128 (37)	5.0
1 month	60 (17)	6.5
2 months	51 (15)	8.1
3 months	72 (21)	8.1
4 months	16 (5)	4.6
≥5 months	17 (5)	3.7
Smear status/grade		
Negative	109 (32)	6.9
Scanty positive	37 (11)	8.4
Positive, grade 1	59 (17)	4.6
Positive, grade 2	51 (15)	6.4
Positive, grade 3+	88 (26)	5.7
Sex		
Female	202 (59)	6.1
Male	142 (41)	6.4
Age category		
≤15	12 (3)	7.8
16–34	171 (50)	5.9
35–54	134 (39)	6.4
≥55	27 (8)	7.9
TB Strain		
LAM4	259 (75)	8.3
Other	85 (25)	0.2

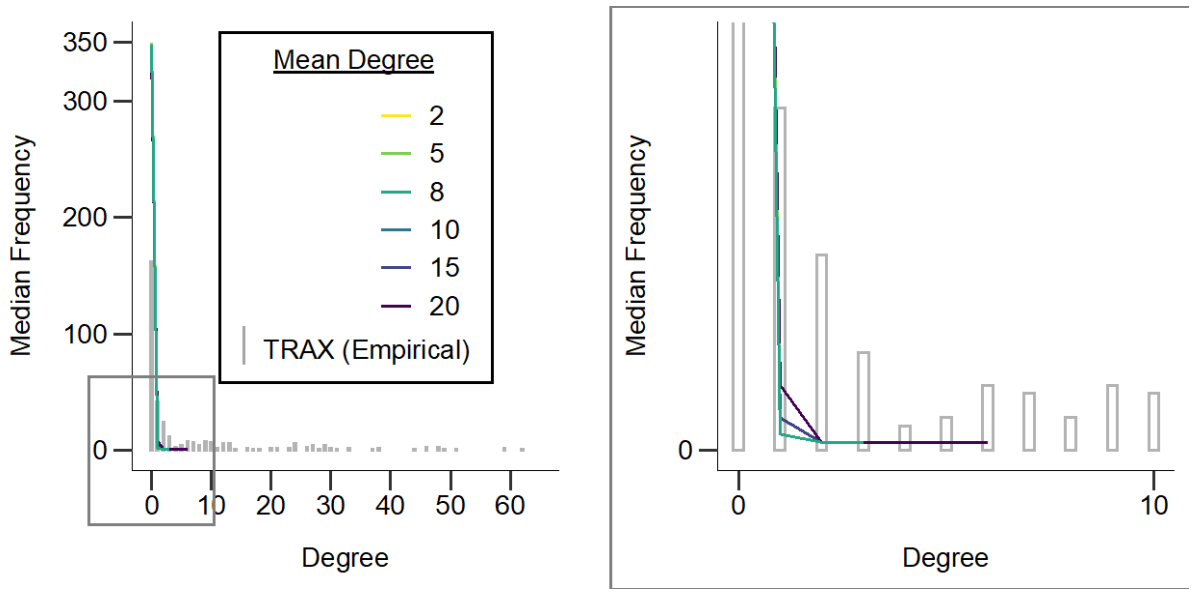
Year		
2011	58 (17)	8.1
2012	107 (31)	5.6
2013	82 (24)	5.8
2014	97 (28)	6.4

Web Figure 2. Mean degree required to reproduce the maximum degree in the empirical network of XDR TB cases, 2011–2014

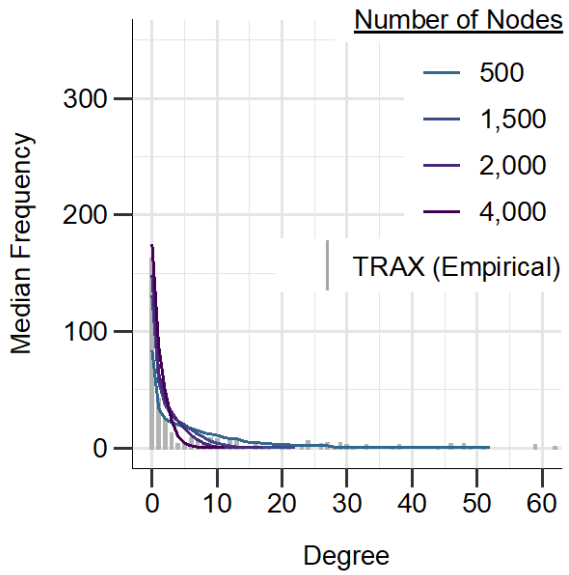


Mean degree required to reproduce the maximum degree in the empirical network. Gray bars show the distribution of the number of links per case, or the degree distribution, of the empirical network (≤5 SNPs) from the TRAX transmission study. Each colored line shows the median degree distribution across 1,000 modeled, sampled networks for the corresponding model. Line color indicates the mean degree, or the average number of transmissions per case, assumed in the complete, simulated network.

Web Figure 3. Degree Distributions of Modeled, Sampled Networks Under Scenarios 1 and 2 Compared to Empirical Network of XDR TB cases, 2011–2014: Figure 2E with zoom (gray box) to show detail.



Web Figure 4. Effect of Modifying Complete Network Size on Modeled, Sampled Networks Under Random Sampling Compared to Empirical Network of XDR TB Cases, 2011–2014.



Degree distributions of empirical (≤5 SNPs) and modeled, sampled networks under different scenarios. Gray bars show the distribution of the number of links per case, or the degree distribution, of the empirical network (≤ 5 SNPs) from the TRAX transmission study; colored lines show the median degree distribution across 1,000 modeled, randomly sampled networks for the corresponding model. Each model makes a different assumption about the total number of XDR TB cases involved in the transmission network during the time period of our transmission study (2011–2014), or the size of the complete transmission network. The model shown has a mean degree in the complete network of 10.

Web Table 9. Effect of modifying network size on modeled, sampled networks compared to the empirical network of XDR TB cases, 2011–2014

Mean Degree	Degree, 10th Percentile of Degree Distribution; Median (IQR)	Degree, 25th Percentile of Degree Distribution; Median (IQR)	Degree, 50th Percentile (**Median**) of Degree Distribution; Median (IQR)	Degree, 75th Percentile of Degree Distribution; Median (IQR)	Degree, 100th Percentile (**Maximum**) of Degree Distribution; Median (IQR)	*P* Value (Median)[4]	*P* Value (IQR)[4]
Target (5-SNP)	**0**	**0**	**1**	**7**	**62**	—	—
Target (3-SNP)	**0**	**0**	**0**	**1**	**22**	—	—
Random sampling (scenario 1) with network size = 4,000							
2	0 (0, 0)	0 (0, 0)	0 (0, 0)	0 (0, 0)	3 (2, 3)	0	(0, 0)
5	0 (0, 0)	0 (0, 0)	0 (0, 0)	1.(1, 1)	4 (4, 5)	0	(0, 0)
10	0 (0, 0)	0 (0,0)	0.5 (0, 1)	1.75 (1, 2)	6 (6, 7)	0	(0, 0)
20	0 (0, 0)	0 (0, 0)	1 (1, 1)	3 (3, 3)	10 (9, 10)	0	(0, 0)
Random sampling (scenario 1) with network size = 2,000 (From Table 3)							
2	0 (0, 0)	0 (0, 0)	0 (0, 0)	0 (0, 1)	4 (4, 5)	0	(0, 0)
5	0 (0, 0)	0 (0, 0)	0 (0, 0)	1.(1, 0)	7 (6, 8)	0	(0, 0)
10	0 (0, 0)	0 (0,0)	1 (1, 1)	3 (3, 3)	11 (10, 12)	0	(0, 0)
20	0 (0, 0)	0 (0, 0)	2 (2, 3)	6 (6, 6)	21 (19, 22)	0	(0, 0.00001)
Random sampling (scenario 1) with network size = 1,500							
2	0 (0, 0)	0 (0, 0)	0 (0, 0)	1 (1, 1)	5 (5, 6)	0	(0, 0)
5	0 (0, 0)	0 (0, 0)	0.5 (0, 1)	2 (2, 2)	9 (8, 10)	0	(0, 0)
10	0 (0, 0)	0 (0, 0)	1 (1, 1)	4 (4, 4)	14 (13, 15)	0	(0, 0)
20	0 (0, 0)	0 (0, 0)	3 (3, 3)	8 (8, 8)	23 (22, 25)	0	(0, 0)
Random sampling (scenario 1) with network size = 500							
2	0 (0, 0)	0 (0, 0)	1 (1, 1)	2 (2, 2)	11 (10, 12)	0	(0, 0)
5	0 (0, 0)	0 (0, 0)	2 (2, 2)	6 (5.75, 6)	22 (21, 24)	0	(0, 0)
10	0 (0, 0)	1 (1, 1)	4 (4, 5)	11 (11, 12)	39 (37, 41)	0	(0, 0)
20	0 (0, 0)	2 (2, 2)	9 (9, 9)	23 (23, 24)	65 (63, 68)	0	(0, 0)

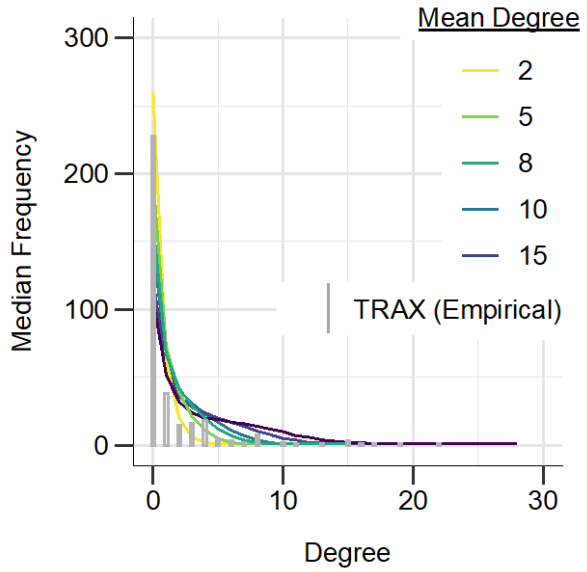
[1] 1,000 networks were simulated from each model, each simulated network was sampled once.
[2] Note that target statistics for both the 5-SNP and 3-SNP empirical networks are shown. These are independent of the results from the modeled networks under scenarios 1 and 2, which are shown in the body of the table.
[3] Median of the 10th percentile of the degree distribution from 1,000 simulated, sampled networks.
[4] *P* values are from a Kolmogorov-Smirnov test with a two-sided alternative hypothesis, calculated using 1,000 bootstrap samples.

Web Table 10. Descriptive Characteristics, Sequencing-Based Network of XDR TB Cases in the TRAX Study (≤3-SNP Threshold), 2011–2014

	No. (%)	Mean
Total network		—
Edges (genomic links)	240	—
Isolates (unlinked cases)	228 (66)	
Overall mean degree	—	1.4
10th percentile	0	
Median degree (IQR)	0 (0,1)	
Maximum degree	22	
Nodes with degree ≥ 10	9 (3)	
By attribute		
HIV status		
HIV-	78 (23)	1.15
HIV+, undetectable VL	133 (39)	1.50
HIV+, detectable VL	133 (39)	1.37
Cough duration		
No cough	128 (37)	1.30
1 month	60 (17)	1.35
2 months	51 (15)	1.75
3 months	72 (21)	1.69
4 months	16 (5)	0.06
5 months	17 (5)	0.71
Smear status/grade		
Negative	109 (32)	1.49
Scanty +	37 (11)	1.73
Positive, grade 1	59 (17)	1.05
Positive, grade 2	51 (15)	1.35
Positive, grade 3+	88 (26)	1.31
Sex		
Female	202 (59)	1.34
Male	142 (41)	1.42

Age category		
<15	12 (3)	1.25
16–34	171 (50)	1.19
35–54	134 (39)	1.53
≥55	27 (8)	1.78
TB strain		
HP	259 (75)	1.79
Other	85 (25)	0.09
Year		
2011	58 (17)	1.84
2012	107 (31)	1.36
2013	82 (24)	1.10
2014	97 (28)	1.34

Web Figure 5. Effect of Reducing SNP Threshold (≤3 SNPs) in the Empirical Network of XDR TB Cases, 2011–2014



Degree distributions of empirical (≤3 SNPs) and modeled, sampled networks under scenario 1. Gray bars show the distribution of the number of links per case, or the degree distribution, of the empirical network (≤3 SNPs) from the TRAX transmission study. Each colored line shows the median degree distribution across 1,000 modeled, randomly sampled networks for the corresponding model. Line color indicates the mean degree, or the average number of transmissions per case, assumed in the complete, modeled network.

References

1. Shah NS, Auld SC, Brust JCM, et al. Transmission of extensively drug-resistant tuberculosis in South Africa. *N Engl J Med.* 2017;376:243–253.
2. Eldholm V, Monteserin J, Rieux A, et al. Four decades of transmission of a multidrug-resistant *Mycobacterium tuberculosis* outbreak strain. *Nat Commun*. 2015;6:7119.
3. Walter KS, Colijn C, Cohen T, et al. Genomic variant identification methods alter *Mycobacterium tuberculosis* transmission inference. *bioRxiv.* 2019:733642.
4. Melsew YA, Gambhir M, Cheng AC, et al. The role of super-spreading events in *Mycobacterium tuberculosis* transmission: evidence from contact tracing. *BMC Infect Dis.* 2019;19(1):244.
5. Ma Y, Horsburgh CR, White LF, Jenkins HE. Quantifying TB transmission: a systematic review of reproduction number and serial interval estimates for tuberculosis. *Epidemiol Infect.* 2018;146(12):1478–1494.
6. Salpeter EE, Salpeter SR. Mathematical model for the epidemiology of tuberculosis, with estimates of the reproductive number and infection-delay function. *Am J Epidemiol.* 1998;147(4):398–406.
7. Ypma RJ, Altes HK, van Soolingen D, Wallinga J, van Ballegooijen WM. A sign of superspreading in tuberculosis: highly skewed distribution of genotypic cluster sizes. *Epidemiology (Cambridge, Mass).* 2013;24(3):395–400.
8. 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.
9. Ismail NA, Mvusi L, Nanoo A, et al. Prevalence of drug-resistant tuberculosis and imputed burden in South Africa: a national and sub-national cross-sectional survey. *Lancet Infect Dis.* 2018;18(7):779–787.
10. Wood R, Middelkoop K, Myer L, et al. Undiagnosed tuberculosis in a community with high HIV prevalence: implications for tuberculosis control. *Am J Respir Crit Care Med.* 2007;175:87–93.
11. Abu-Raddad LJ, Sabatelli L, Achterberg JT, et al. Epidemiological benefits of more-effective tuberculosis vaccines, drugs, and diagnostics. *Proc Natl Acad Sci U S A.* 2009;106(33):13980.
12. Naidoo CC, Pillay M. Increased in vitro fitness of multi- and extensively drug-resistant F15/LAM4/KZN strains of *Mycobacterium tuberculosis*. *Clin Microbiol Infect.* 2014;20:O361–O369.
13. Cohen KA, Abeel T, Manson McGuire A, et al. Evolution of extensively drug-resistant tuberculosis over four decades: whole genome sequencing and dating analysis of *Mycobacterium tuberculosis* isolates from KwaZulu-Natal. *PLoS Med.* 2015;12:e1001880.
14. Wood R, Racow K, Bekker L-G, et al. Indoor social networks in a South African township: potential contribution of location to tuberculosis transmission. *PloS One.* 2012;7(6):e39246–e39246.
15. Hall P, Härdle W, Simar L. On the inconsistency of bootstrap distribution estimators. *Comput Stat Data Anal.* 1993;16(1):11–18.