UC Davis
UC Davis Previously Published Works

Title

Strong effect of demographic changes on Tuberculosis susceptibility in South Africa

Permalink

https://escholarship.org/uc/item/1rv9p76r

Journal

PLOS Global Public Health, 4(7)

ISSN

2767-3375

Authors

Oyageshio, Oshiomah P

Myrick, Justin W

Saayman, Jamie

et al.

Publication Date

2024

DOI

10.1371/journal.pgph.0002643

Copyright Information

Peer reviewed

Strong effect of demographic changes on Tuberculosis susceptibility in South Africa

Oshiomah P. Oyageshio[1]☯*, Justin W. Myrick[2]☯, Jamie Saayman[3], Lena van der Westhuizen[3], Dana R. Al-Hindi[4], Austin W. Reynolds[5], Noah Zaitlen[6], Eileen G. Hoal[3], Caitlin Uren[3,7], Marlo Möller[3,7]*, Brenna M. Henn[1,2,4]*

1 Center for Population Biology, University of California, Davis, Davis, California, United States of America, 2 UC Davis Genome Center, University of California, Davis, Davis, California, United States of America, 3 DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa, 4 Department of Anthropology, University of California, Davis, Davis, California, United States of America, 5 Department of Microbiology, Immunology, and Genetics, School of Biomedical Sciences, University of North Texas Health Science Center, Fort Worth, Texas, United States of America, 6 Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, California, United States of America, 7 Centre for Bioinformatics and Computational Biology, Stellenbosch University, Stellenbosch, South Africa

☯ These authors contributed equally to this work.
* oyageshio@ucdavis.edu (OPO); marlom@sun.ac.za (MM); bmhenn@ucdavis.edu (BMH)

Abstract

South Africa is among the world's top eight tuberculosis (TB) burden countries, and despite a focus on HIV-TB co-infection, most of the population living with TB are not HIV co-infected. The disease is endemic across the country, with 80–90% exposure by adulthood. We investigated epidemiological risk factors for (TB) in the Northern Cape Province, South Africa: an understudied TB endemic region with extreme TB incidence (926/100,000). We leveraged the population's high TB incidence and community transmission to design a case-control study with similar mechanisms of exposure between the groups. We recruited 1,126 participants with suspected TB from 12 community health clinics and generated a cohort of 774 individuals (cases = 374, controls = 400) after implementing our enrollment criteria. All participants were GeneXpert Ultra tested for active TB by a local clinic. We assessed important risk factors for active TB using logistic regression and random forest modeling. We find that factors commonly identified in other global populations tend to replicate in our study, e.g. male gender and residence in a town had significant effects on TB risk (OR: 3.02 [95% CI: 2.30–4.71]; OR: 3.20 [95% CI: 2.26–4.55]). We also tested for demographic factors that may uniquely reflect historical changes in health conditions in South Africa. We find that socioeconomic status (SES) significantly interacts with an individual's age ($p = 0.0005$) indicating that protective effect of higher SES changed across age cohorts. We further find that being born in a rural area and moving to a town strongly increases TB risk, while town birthplace and current rural residence is protective. These interaction effects reflect rapid demographic changes, specifically SES over recent generations and mobility, in South Africa. Our models show that such risk factors combined explain 19–21% of the variance (r^2) in TB case/control status.

Introduction

Tuberculosis (TB) is the world's leading cause of death due to infectious disease, currently greater than COVID-19 [1]. The causative agent, *Mycobacterium tuberculosis (M.tb)*, is an obligate intracellular pathogen mainly infecting the lungs, and sometimes other organs [2, 3]. Approximately 25% of the world's population is infected with *M.tb* and the annual death toll is similar to COVID-19 (~1.5 million deaths). South Africa is amongst the top 30 'high burden' countries coping with TB, TB/HIV co-infection, and multi-drug resistance or rifampicin-resistant TB (MDR/RR-TB). TB is South Africa's leading natural cause of death [4] with an extremely high prevalence (446/100,000, [5]) and accounts for 3.3% of all global TB cases [1]. The Northern Cape presently has the highest TB incidence in South Africa (ZF Mgcawu district 926/100,000 [5]), but the lowest HIV prevalence (7.1% vs 13.5% National average), including the lowest density of people living with HIV [6].

Determinants of active TB progression are multifaceted, including: genetics, nutrition, social and economic conditions, behavior, and sex-specific biology [1, 7, 8]. Initial *M.tb* infection is largely determined by exogenous factors, such as TB prevalence in the community, population density (e.g., prisons), and working conditions (e.g., mining, healthcare workers) [9–13]. The lifetime risk of progressing to active TB following infection is 10%. This risk is the highest within the first 5 years of initial infection and is typically considered to be mediated by the host's innate and cell-mediated immune system [9, 14]. Individual (i.e., host) factors, however, have also been shown to increase risk of progressing to active disease. These include HIV/AIDS, poor nutrition or low body mass index, indoor air pollution (e.g., cooking with wood and poor ventilation, smoking, alcohol abuse, diabetes mellitus, and intravenous drug use [1, 9, 14, 15]. Studies in India have shown undernutrition to be among the strongest determinants for TB risk [15]. In South Africa specifically, poor living conditions, unemployment, low SES, age and male gender, race, smoking, and marital status have all been identified as contributing to TB risk [16–21].

The extent of these determinants' effects can vary across and within populations, necessitating epidemiological studies in differing contexts and communities [8]. Compared to medium or high-TB-incidence countries, the effect sizes for alcohol abuse, homelessness, and intravenous drug use are stronger in low-incidence populations [22]. In South Africa, multilevel modeling approaches have shown that provincial [16] and community income inequality [18] have strong effects on TB incidence and progression, independent of individual-level risk factors.

HIV increases TB risk by 20-fold, the largest known risk factor for progression to active TB, and TB is the leading cause of AIDS-related deaths [23]. The effect of HIV on suppressing the host immune system can reactivate a latent *M.tb* infection and increase susceptibility to initial infection [14, 23, 24]. Despite HIV being the strongest TB determinant, other TB risk factors explain the majority of global TB cases [9]. In South Africa, 59% of people with TB on the National TB Programme (i.e. on TB medication) are co-infected with HIV [25]. However, South Africa's first national TB prevalence survey found that only 28% of people with TB were also people living with HIV (PLWHIV) [25], a finding underscoring the necessity to extend TB research to those living without HIV in high TB burden areas. At the provincial level in South Africa, HIV prevalence explains little of TB incidence ($r^2 = 0.036$) [26]. The Western and Northern Cape Provinces have among the highest TB incidence yet the lowest HIV prevalence [27].

Here, we present a TB case-control study characterizing the individual-level risk factors for TB progression among HIV-negative patients with suspected TB from the Northern Cape. The Northern Cape has the highest TB incidence but the lowest HIV prevalence and PLWHIV

density, and overall low population density, canonical risk factors do not appear to be driving the extraordinary incidence rates. To focus on factors other than immune suppression, we exclude PLWHIV from the analysis. Controls from our study sample are people with suspected TB from local health clinics who were microbiologically confirmed to be negative for active TB. Controls are assumed to either have been previously exposed to or infected with *M.tb* (i.e., LTBI). In South Africa, TB transmission is driven largely by community spread, rather than household contacts [28, 29]. Cases, in contrast, are people who have microbiologically confirmed active TB or self-report a past active TB episode. We test three separate models comprising common risk factors, as well as factors that may uniquely affect South Africa. We find that exogenous factors like SES, cohort age, and residence/birthplace have a strong effect on TB progression, often equal to or greater than endogenous factors like gender or smoking/alcohol. These results suggest further research into the causal mechanisms behind exogenous risk factors and opportunities for TB prevention are warranted.

Methods

Research ethics statement

This study has been approved by the Health Research Ethics Committee (HREC) of Stellenbosch University (N11/07/210A) and the Northern Cape Department of Health (NC2015/008). All participants were adults (18 years and older) and provided written informed formal consent. Authors Justin W. Myrick, Jamie Saayman, Lena van der Westhuizen and Marlo Möller had access to identifiable information about participants as they were directly involved in data collection or database management. Access to these records commenced on 26th January 2016, and is still ongoing as it is an integral part of the Northern Cape Tuberculosis Project (NCTB).

Inclusivity in global research

Additional information regarding the ethical, cultural, and scientific considerations specific to inclusivity in global research is included in the Supporting Information (S1 Checklist).

Study design and recruitment

Participants (18 years and older) provided written informed consent and were recruited from 12 community health clinics from the ZF Mgcawu district in the Northern Cape Province of South Africa from 26th January 2016 – 15th May 2017, and 11th December 2018 – 11th March 2020. Community health clinics are the front line for TB screening and treatment, visited by 87% of people who seek TB care [25]. TB nurses referred patients with suspected TB (with ≥ 2 TB symptoms: cough for ≥ 2 weeks, night sweats, weight loss, and fever ≥ 2 weeks, or interaction with a TB index contact) and known TB patients to our on-site RAs. All study participants took a clinic-administered sputum GeneXpert Ultra test for active TB at the time of the study interview and provided saliva for genotyping. Clinic medical charts were accessed by a staff research nurse to record GeneXpert test results and verify HIV status and TB history.

Case-control assignment

Cases and controls were assigned based on both the participant's medical charts and self-reported data (Fig 1). Cases include anyone with active pulmonary TB in their lifetime *and* that was HIV-negative. Thus, cases could be partitioned into 1) clinically confirmed active TB (n = 208) at the time of enrollment, and 2) self-reported past TB episode(s) (n = 166). GeneXpert results, diagnostic test date, TB strain (drug resistance), and TB medication regimen were
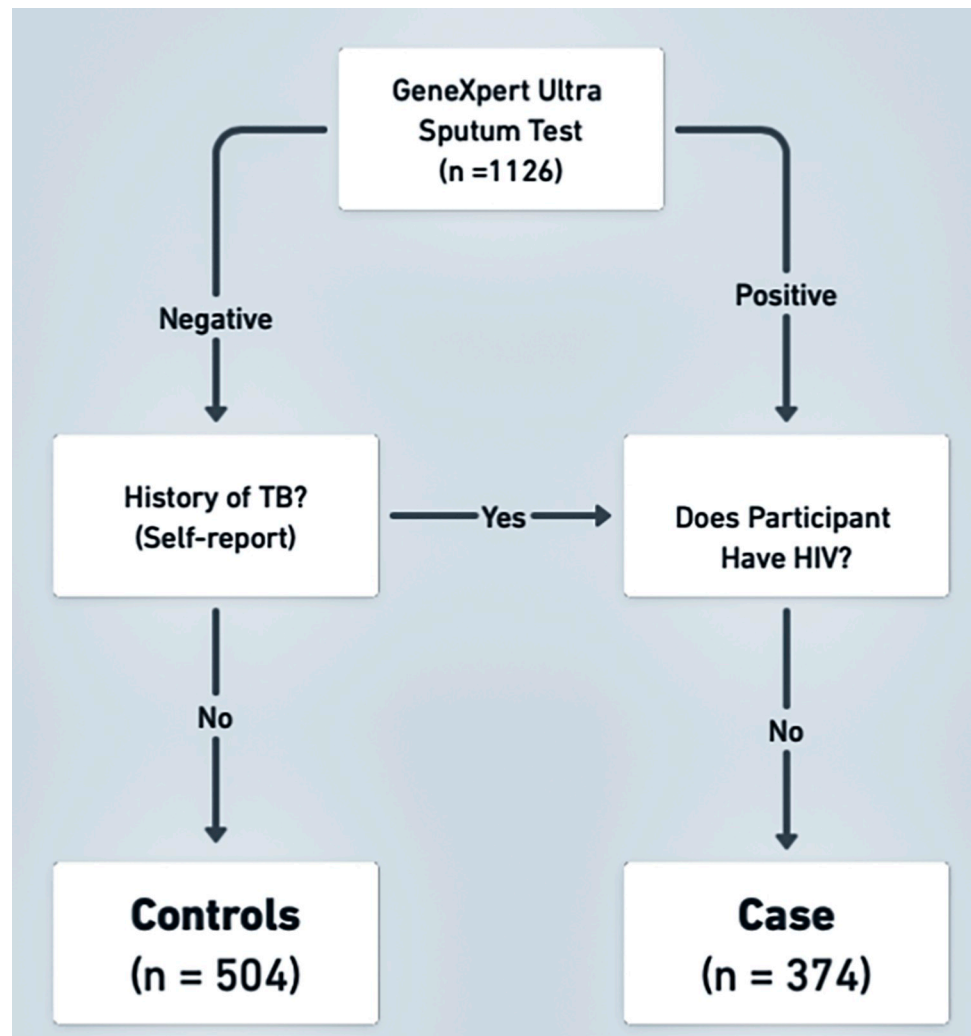
Fig 1. Case-control decision tree. Study participants were categorized as cases or controls based on medical record information and self-reported data. All participants were GeneXpert tested for active TB infection at the time of enrollment. Past TB episodes were self-reported and cross-referenced with medical records when available.

https://doi.org/10.1371/journal.pgph.0002643.g001

used to validate clinically confirmed progression to active TB. Past TB episodes are based on self-report, mainly due to older medical charts which were not reliably available, discarded, or difficult to locate by clinic staff.

We defined controls as HIV-negative clinic patients with suspected TB symptoms who had a negative GeneXpert Ultra result, and no history of active pulmonary TB at the time of study enrollment. Controls in our study are likely to be latently infected with *M.tb* (LTBI) or to have been exposed in their lifetime. A majority of the population in high TB burden South African suburbs are LTBI, 88% by ages 31–35 [30].

Our exclusion criteria removed participants with unknown TB or HIV status, as well as PLWHIV.

Study covariates

We collected demographic information that included date of birth, place of birth, current residence, self-identified gender, self-reported ethnic identity, and parental ethnic identities.

Behavioral variables include smoking, drinking and diabetes (Supplementary Methods in S1 Text). In our analyses, we only used binary measures for smoking, drinking and diabetes ("Do you smoke?", "Do you have diabetes?"). Age at enrollment was used as a continuous variable for all analyses and binned for calculating empirical odds. Socioeconomic status "SES" was operationalized as number of years of education, i.e., the highest completed level of education. McKenzie et al. [31] have shown education level, in this dataset, positively predicts body mass index in TB controls, tracking access to resources and food security if only a crude measure.

Residence and birthplace locations are categorized as rural (≤2000 people) and town (>2,000 people). Population size was derived from the South African census and when census data was absent, e.g., a farm, we used Google Earth (earth.google.com) to estimate population size based on the number of dwellings. Places that did not have a census size available in Stats SA (statssa.gov.za) typically were very small communities, like a farm or a small settlement. By visualizing the settlement through Google Earth, we were able to estimate whether the community size was >2000 estimated people. We used the average household size from the census data from all locations in the district listed by Stats SA (statssa.gov.za), then counted the number of dwellings in blocks and multiplied by the average household size [4] to get an estimated population size. While an imprecise measure of population size, the lack of government census data for a community is itself an indicator of its rural locality.

Statistical analyses

Statistical analyses were performed in R (version 4.2.3). We calculated Pearson correlations with the R package *ggcorrplot*. All categorical variables were numerically coded to "0" and "1". Classification models for our binary, qualitative dependent variable ("case"/ "control") included logistic regression and random forest—a machine learning classifier robust to non-linear associations and unknown variable interactions [32] (Supplementary Methods in S1 Text). To calculate empirical odds, we binned our participants into 7 age groups, dividing the number of controls by the number of cases in each age bin. All R scripts for analyses are available at https://github.com/oshiomah1/NCTB-Epidemiology-Project.

Imputing missing data

To maximize our sample size, we imputed missing data for diabetes, smoking, and years of education (Table C in S1 Text). The proportion of missing data was overall low, i.e. below 5%, except for alcohol use at for which missingness was 35%; the alcohol use measure was implemented after a pilot study. We chose to exclude alcohol from imputation and model analyses due to high missingness. Multivariate imputation was performed using the R package MICE, implementing chained equations in which every variable with incomplete data is imputed conditional on all the data from other variables in the dataset [33]. To initiate the MICE procedure, we created a matrix of variables consisting of age, gender, height, HIV, mother's ethnicity, father's ethnicity, diabetes, smoking status, and years of education. Notably, ethnicity and height variables were not used in our regression analysis but bore potential relevance to our missing variables, so they were included to improve the statistical inference of our imputation. We set the parameters of our imputation using recommended settings [33, 34], generating two imputed datasets ($m = 2$) that were run for 10 iterations each (*maxit = 10*). We used a classification and regression tree method which is robust in epidemiological datasets similar to ours [35].

To cross-validate our imputation method, we randomly sampled ten percent of known values in each variable and converted them to missing values (Table C in S1 Text). Next, these missing values were imputed using the procedure described above, and the missing value was

compared to the original value. For continuous variables, the average percent difference between imputed and original value was used to calculate the cross-validation (CV) score while the average accuracy of the imputed variable was used to generate the CV score for binary variables. This procedure was carried out on one variable at a time for 100 iterations. Cross-validation results revealed that years of education and diabetes were sufficiently imputed (CV score > 10%).

Obtaining and visualizing model coefficients

After MICE imputation, we used the 'psfmi' package [36] to implement our logistic regression models, obtaining pooled odds ratios using Rubin's Rules [37]. Each model was Bonferroni corrected using a baseline of $p<0.05$. For the Residence model, we set lifetime rural dwellers as the baseline and manually calculated contrasts for the other three comparisons. To illustrate the covariate effects from our models, we extracted the first imputed dataset from the MICE output, used the R 'glm' function to implement the logistic regression models, and then used the 'effects' package [38] to visualize the odds of Active TB.

Genetic data processing & ancestry estimation

A subset of participants ($n = 159$) was genotyped for >2 million single nucleotide polymorphisms (SNPs) on the Illumina H3Africa array. Genetic data processing involved DNA extraction from saliva samples, common variant calling with GenomeStudio, rare variant calling with zCall, and further data cleaning using *plink2* (Supplementary Methods in S1 Text). Global [i.e. genome-wide] ancestry estimates were calculated using ADMIXTURE v1.13 [39]. The Luhya, Maasai, Himba, British, Palestinian, Chinese, Bangladeshi, Tamil, Ju|'hoansi San, Khomani San, and Nama populations were used as reference groups encompassing all major ancestry sources. ADMIXTURE was run in groups of maximally unrelated individuals to avoid biasing the ancestry estimates. We assumed k = 5 possible ancestries, inferred in unsupervised mode for each of the running groups. After matching clusters, we merged ancestry estimates across all running groups, averaging individuals that appeared in multiple running groups using pong [40]. We further tested whether population stratification affected the results of the logistic regression models by including 10 principal components (computed with *plink2*) and re-computing regressions for just the subset of $n = 159$ individuals (Supplementary Methods in S1 Text).

Results

TB case-control classification

1,126 participants were partitioned into preliminary cases, preliminary controls, and unverified TB status (571, 504, and 51 respectively). After excluding, participants with unverified TB status and either unknown or positive HIV status, 774 participants remained in the study (374 cases and 400 controls; Table A in S1 Text).

Socio-behavioral and demographic characteristics of the cohort

Men and women were equally represented in the dataset (384 vs. 390, respectively, Table A in S1 Text). Men were more likely to drink alcohol (p < 0.001) and smoke (p < 0.05). A high fraction of our participants self-reported smoking (67%) and drinking alcohol (46%); smoking and drinking were moderately correlated with each other ($r = 0.36$, p < 0.05; Fig A in S1 Text). Women were more likely to have diabetes (p = 0.0004; Fig A in S1 Text) and, on average, had more education than men (female mean = 8.2 years, male mean = 7.8 years).
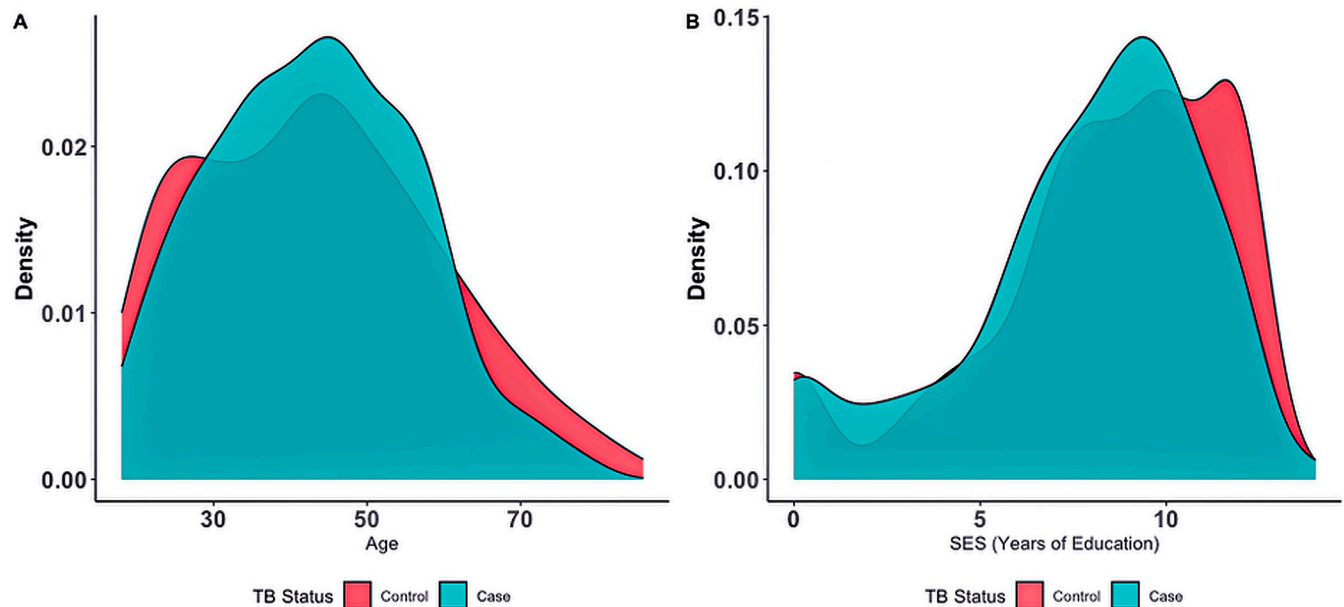
Fig 2. Density plots of continuous variables A) Age by case-control status B) SES by case-control status.

We use the number of years of education as a proxy for socio-economic status "SES" (*Methods*). The mean educational attainment in our cohort is 8 years, equivalent to completing primary school, and is similar between individuals recruited in rural areas and towns [ANOVA, $p > 0.1$]. In the ZF Mgcawu District census [41], 13% of adults have not completed primary school as compared to 25.3% of our participants. Age was moderately correlated with SES ($r = -0.5$, $p < 0.05$; Fig A in S1 Text) such that older participants tended to have lower SES. Cases and controls had similar mean ages, 43.6 and 43.1 years respectively (Wilcoxon rank sum *p*-value = 0.959) (Fig 2A). We found a significant difference in SES between cases and controls, with mean of 7.7 years and 8.3 years, respectively (Wilcoxon rank sum *p*-value = 0.0019) (Fig 2B).

To investigate the possibility of selection bias we computed empirical odds of active TB by age group. Assuming that age is a cumulative measure of exposure (that is, capturing the amount of time someone is exposed to TB), the empirical odds of TB should increase monotonically with age. We observe a non-monotonic trend where the odds of active TB progressively increase from ages 18 up to 38, then reverses, progressively decreasing starting at age 39 up till the 79–88 age group having the lowest empirical odds (Fig B in S1 Text).

Ethnicity and genetic ancestry

Individuals were asked to self-identify their ethnicity without categorical prompts. 88.4% of participants [both TB cases and controls] self-identify as Coloured, followed by 4.2% as a Khoe-San ethnicity (e.g., Nama, San), 4.6% as Tswana, 1.3% as Xhosa, and 1.9% as "other". Whilst we acknowledge that in some contexts the term "Coloured" has derogatory connotations, in South Africa it is a recognized ethnicity as well as a racial category. People who self-identify using this term tend to have genetic ancestries from multiple geographic origins, including the indigenous Khoe-San groups (e.g., Khoekhoe, San), Bantu-speaking, European, Indian, Malaysian (Southeast Asian) slaves, or people of mixed ancestry and their descendants [42]. The use of "Coloured" in this context reflects the self-identified cultural attributes of the
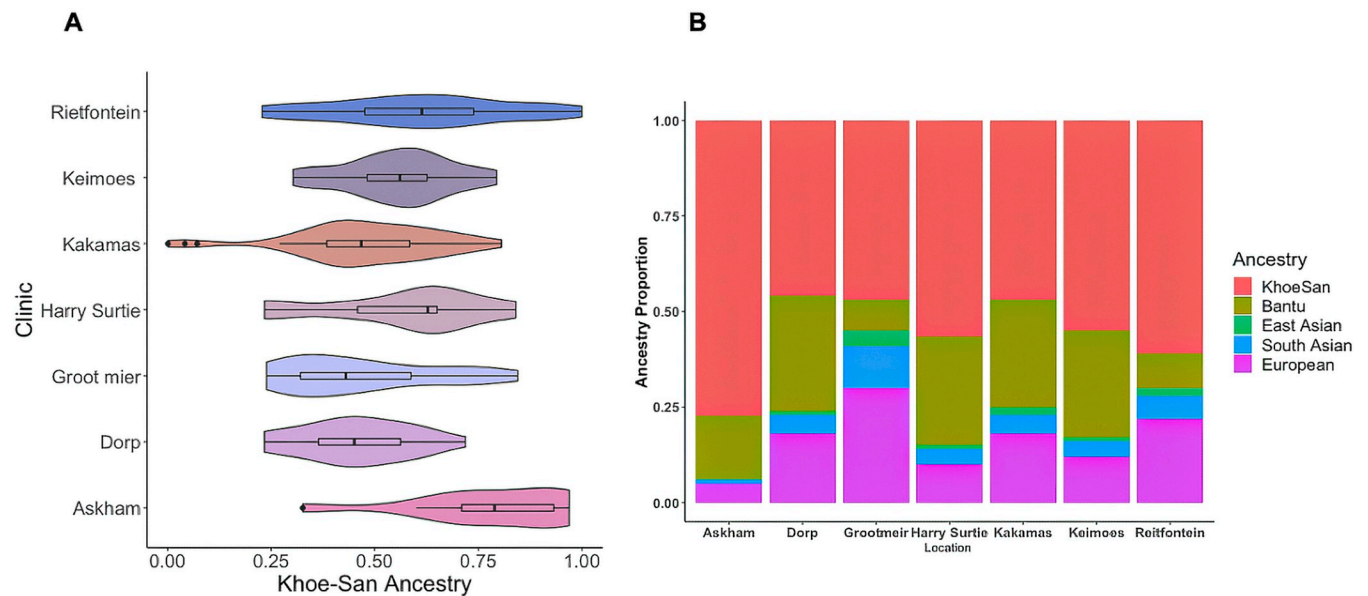
Fig 3. Khoe-san ancestry is the primary genetic ancestry in clinics from the northern cape, South Africa. A subset of participants (n = 159) was genotyped to obtain the average genome-wide ancestry proportions across all individuals for each clinic. A) Khoe-San ancestry is the largest proportion of ancestry in our sample, it varies significantly across study sites. The boxplots show the median, the 25th and 75th percentile and 1.5 times said percentile, and all outliers as dots. B) The study population is admixed with 5 distinct ancestries with the southern African indigenous Khoe-San ancestry being the highest proportion of ancestry at all study sites.

participants, as well as possible historical and genetic attributes. Ethnicity is reported out of respect for participants' choice of identity.

Genetic ancestry characterization was performed for 159 participants to assess if there was significant variation among clinics which could potentially confound analysis. Since genetic information was not available for all participants in the study, individual ancestry was not directly factored into the logistic regression models and random forest models. Khoe-San ancestry varied across clinic locations (Fig 3A) but remained the majority ancestry at each site (mean = 56%), followed by Bantu-speaking African ancestry (mean = 21%), European ancestry (mean = 16%), South Asian ancestry (mean = 5%), and East Asian ancestry (mean = 2%) (Fig 3B).

There was statistically significant variation in ancestry at the clinic level (ANOVA: Khoe-San, F = 6.9, p < 1e-06; Bantu-speaking, F = 7, p < 6e-05; European, F = 5, p < 6e-05; South Asian, F = 8, p < 4.8e-08; East Asian, F = 4.5, p < 0.001, Fig 3A). The statistically significant differences in the proportion of mean ancestry were generally between clinics in the Kalahari (Askham and Rietfontein, except Groot Mier) versus clinics along the Orange River (Harry Surtie, Dorp, Kakamas, and Keimoes) determined by Tukey HSD post hoc tests. The Kalahari clinics tended to have more Khoe-San ancestry (20–30% more) than the Orange River clinics and less Bantu-speaking African ancestry (~20% less). Notably, Groot Mier in the Kalahari had more European (11–19% more) and greater South and East Asian ancestry (4–10% and 2–3% more, respectively) than most other clinics (Fig 3B), likely due to Groot Mier's history as an early European colonial post [43].

Hypotheses and models of progression to active TB

We designed three logistic regression models [44] to examine the risk factors that determine TB case/control status in our sample. Our first model, which we termed the "common risk

factor" model (n = 774), includes six covariates known to be common behavioral or demographic risk factors for TB. We hypothesized that risk factors identified from prior studies are also significantly associated with active TB progression in our HIV-negative population.

Common risk factor model: TB Status ~ gender + smoking + diabetes + residence + age + SES

Health disparities are one of the many consequences of apartheid in South Africa [45, 46]. The end of apartheid circa 1994 improved social mobility and educational access; however, health disparities in the Northern Cape (and other provinces) still remain problematic [47]. We formulated two alternative models which included variables potentially important to South African populations, involving the change in SES over time, and migration between rural and urban areas. We hypothesized that there are differential effects of SES (education) on TB status due to the sociopolitical effects of apartheid. To capture the effect of lived experience vis-à-vis Apartheid on TB outcomes we designed an "SES model" (n = 774). This model includes the common risk factors as above but allows for an interaction between age and SES to account changing economic conditions over the past eighty years. For example, completion of a high school equivalent education in 1960 did not afford the same economic benefits as completion of a high school education in 2010. We predict that for younger cohorts, higher SES is protective against TB; in contrast, for individuals born during apartheid, higher SES would have little effect on lifetime TB status. Age was kept as a continuous variable because apartheid was not a historically binary event.

SES model: TB Status ~ common risk factor model + age * SES

Residing in an urban or rural environment is an established risk factor for TB status [11, 48–50] and was included in our common risk factor model, as above. Previously, we have shown that migration from an individual's natal town has increased over the past two generations in the rural Northern Cape Province [31]. In addition, a recent longitudinal study leveraging data from South Africa's National Health Laboratory Service showed that incorporating cross-municipality migration improves the ability to predict TB incidence [51]. For our "residence model", we hypothesized that migrating from a rural to urban area in one's lifetime increases the odds of active TB status due to greater exposure to *M.tb*. Here, we include an interaction between current residence and birthplace in the common risk factor model. Setting this interaction allows us to examine four patterns, namely: lifetime rural residence, rural birthplace to urban residence, urban birthplace to rural residence, and lifetime urban residence.

Residence Model: TB Status ~ common risk factor model + residence * birthplace

The common risk factor model (pseudo R^2 = 19%, n = 774, Table 1) performed slightly worse than the SES model (pseudo R^2 = 21%, n = 774, Table 2). The residence model had a smaller sample size (pseudo R^2 = 19%, n = 720, Table 2) than the common risk factor model due to missing birthplace data for some individuals. For an equal comparison, we re-ran the common risk factor and SES models with same individuals as in the Residence model. For the reduced dataset the Residence model and SES model had comparable pseudo R^2 while the common risk factor model had a slightly worse value (pseudo R^2 = 18% vs. 19%) (Table B in S1 Text). Therefore, we present results from all three models and contrast the variable effects (Tables 1 and 2). All significance levels were Bonferroni corrected, assuming an α = 0.05.

Common risk factors

Across the three models, males consistently have three times the odds of active TB than females (OR = 3.01 [2.20,4.12], $p < 0.001$; Tables 1 and 2, and Fig C in S1 Text). All logistic regression models showed insufficient statistical evidence for smoking (common risk factor

Table 1. Odds ratios and *p*-values for the demographic and socio-behavioral variables used in the common risk factor model.

	Common Risk Factor (n = 774)	
	Odds Ratio (95% CI)	*p* value
(Intercept)	0.23 [0.08, 0.66]	0.007
Residence	3.20 [2.26, 4.55]	p < 0.001**
Gender	3.01 [2.20, 4.12]	p < 0.001**
Age	0.99 [0.97, 1.00]	0.046
Years of education (SES)	0.91 [0.86, 0.97]	0.006*
Smoker	1.32 [0.94, 1.85]	0.111
Diabetes	1.27 [0.67, 2.43]	0.462
n	774	
R²	0.19	

* alpha = 0.05

** alpha = 0.01

model: OR = 1.32 [0.94, 1.85], *p* = 0.11; Tables 1 and 2), and diabetes (common risk factor model: OR = 1.27 [0.67, 2.43], *p* = 0.46; Tables 1 and 2) on TB risk. Despite the lack of significance, we note that smoking had an effect size in the expected direction (Fig C in S1 Text). The variable with the strongest effect size was current residence in towns–areas with a population size greater than 2,000 peoples (OR = 3.20 [2.26, 4.55], *p* <0.0001; Tables 1 and 2 and Fig C in S1 Text).

Age interacts with SES

In the common risk factor model, age does not significantly affect TB risk, but higher SES has a protective effect (OR = 0.91 [0.86, 0.97], *p* = 0.006; Table 1). In the SES model, SES significantly affects TB status depending on age group (OR = 1.01 [1.00, 1.01], *p*< 0.001, Table 2).

Table 2. Odds ratios and *p*-values for the demographic and socio-behavioral variables used in the residence model and SES model.

	Residence Model		SES Model	
	Odds Ratio (95% CI)	*p* value	Odds Ratio (95% CI)	*p* value
(Intercept)	0.36 [0.12, 1.12]	0.077	3.69 [0.50, 27.36]	0.197
Birthplace^	2.84 [1.28, 6.21]	0.005*		
Residence^	6.10 [2.97, 12.5]	p <0.001**	3.29 [2.30, 4.71]	p <0.001**
Gender^	3.04 [2.19, 4.21]	p <0.001**	3.02 [2.20, 4.15]	p <0.001**
Age	0.98 [0.97, 1.00]	0.022	0.94 [0.91, 0.97]	*p* <0.001**
Years of education (SES)	0.92 [0.86, 0.98]	0.013	0.66 [0.54, 0.81]	*p* <0.001**
Smoker	1.32 [0.93, 1.89]	0.124	1.22 [0.85, 1.73]	0.274
Diabetes	1.49 [0.74, 3.00]	0.267	1.28 [0.67, 2.45]	0.453
Years of education: Age			1.01 [1.00, 1.01]	*p* <0.001**
Birthplace: Residence	0.33 [0.13–0.86]	0.024		
n	720		774	
R²	0.19		0.21	

* alpha = 0.05

** alpha = 0.01

^ We report the OR with a baseline of rural = 1, for Birthplace; a baseline of town = 1, for Residence; for Gender the OR is listed for males, and for Smoking and Diabetes the OR is listed for a positive response.
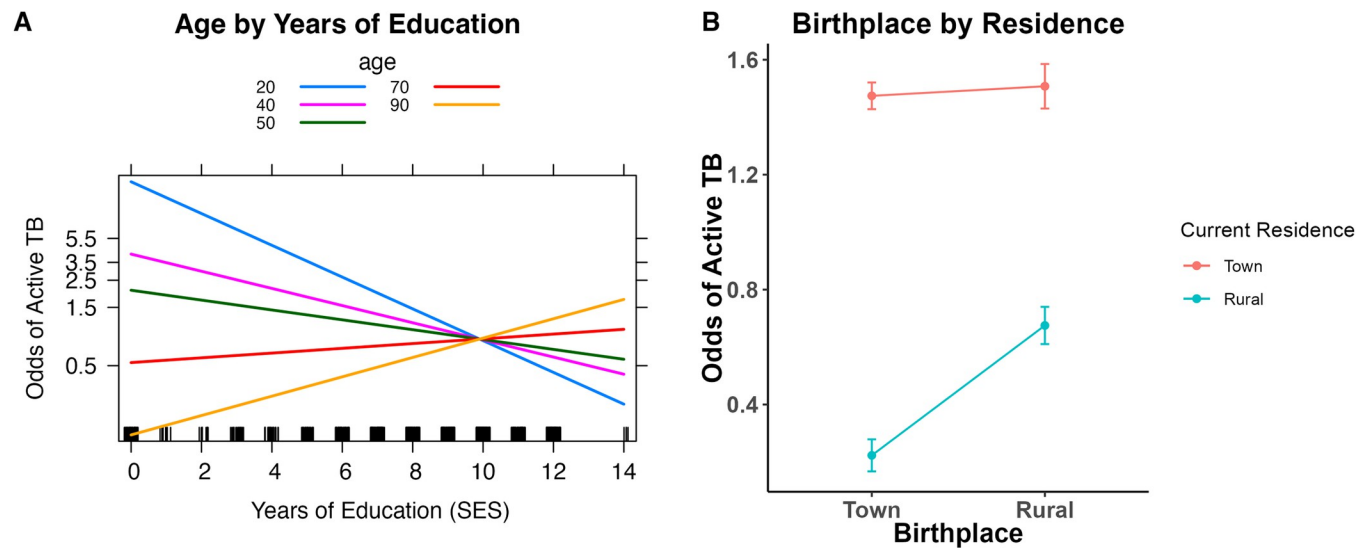
Fig 4. Logistic regression interaction plots. A) The odds of active TB by education level vary across age groups (shown above by the different color lines). More years of education decreases the odds of active TB in younger age groups, but this pattern reverses in the oldest age groups. B) Effect plot from the residence model visualizing an interaction term between birthplace residence and current residence. Regardless of birthplace, the odds of active TB is highest in individuals who currently reside in towns. Individuals born in towns and currently residing in rural areas have the lowest odds of active TB.

The effect is such that higher SES at younger ages (18–59 years old) is protective against TB, and higher SES at older ages (>59 years) increases risk (Fig 4)

TB risk by residence and birthplace

We analyzed the relationship between TB status and a change of locality between birthplace (rural area or town) and current residence (rural area or town) during an individual's lifetime. We expected to see a difference in the odds of active TB between lifelong residents and those who have moved between locales. Under such a model, lifelong rural dwellers would have the lowest odds and lifelong town dwellers would have the highest odds. We set an interaction term between current residence and birthplace classified into town/rural, (OR = 0.33 [0.13–0.86], p = 0.024, Table 2). To break down the interaction effect, we set lifelong residents of rural areas as the baseline, comparing the three other residence patterns with this baseline group. We found that lifelong town dwellers had about twice the odds of active TB (OR = 2.16 [1.43–3.28], $p < 0.001$) relative to the baseline. Individuals born in a rural area and currently residing in a town had similar outcomes as lifelong town dwellers (OR = 2.19 [1.14–4.20], $p = 0.018$). Taken together, these show that town residence increases risk regardless of birthplace. Interestingly, individuals who were born in a town and later moved to rural areas are even more protected than individuals born and currently residing in rural areas (OR = 0.33 [0.16–0.71], $p = 0.004$) (Fig 4).

Random forest modeling

As an alternative to logistic regression, we trained a random forest model to classify TB status utilizing all the variables from the common risk factor model. We configured the model to grow 5000 classification trees (Supplementary Methods in S1 Text). The model assigned gender, current residence, SES, and age as the top important independent variables (Fig 5). Diabetes and smoking were classified as uninformative predictors for TB status. The model had an overall "out-of-bag" misclassification rate of 23%.
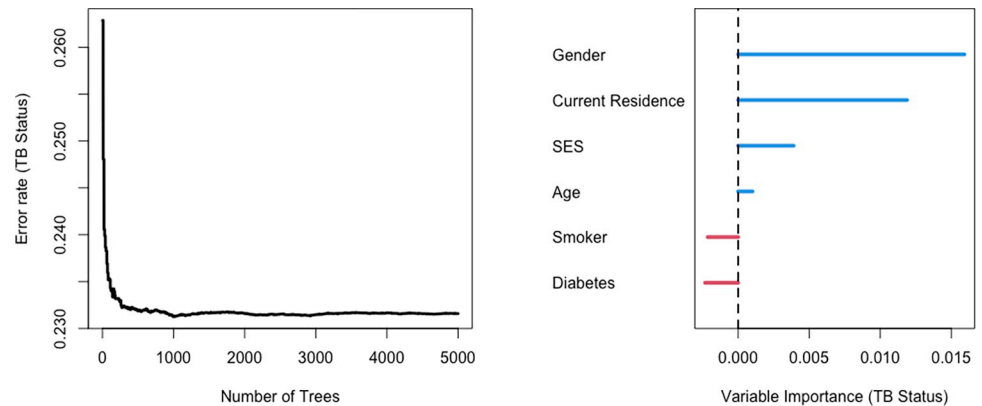
Fig 5. Random forest model with common risk factor variables (n = 774). Random subsets of all 6 variables on the y-axis were used to grow 5000 trees to classify participants into cases and controls. The model had an overall "out-of-bag" misclassification rate of 23%. Variables with higher variable importance are most crucial for case-control classification. Predictor variables with negative variable importance values worsen the ability of the model to classify TB status.

https://doi.org/10.1371/journal.pgph.0002643.g005

Discussion

We examined common demographic risk factors for TB, constructing the largest TB epidemiological study in a Northern Cape clinical population (n = 774), to our knowledge. We show that gender, SES, and current residence locality are significant variables or important TB risk predictors, using logistic regression and random forest models. Neither smoking nor diabetes is associated with increased TB risk in any model. Among the three logistic regression models, interacting SES by age ("SES model"), and birthplace by residence ("residence model"), had similar explanatory power, improving on the common risk factor model.

In South African townships, *M.tb* is community spread [28, 29] and about 88% of adults are infected by ages 31–35 [30]. Here, we demonstrate the utility of sampling in high disease incidence populations to rapidly build datasets with large sample sizes of TB cases and controls with similar pathogen exposure. Validating *M.tb* exposure in controls is traditionally done by a tuberculin skin test (TST) and/or interferon-gamma release assay (IGRA; e.g., Quanti-FERON), thereby assigning LTBI status. We did not perform IGRA and TST tests for all controls in our cohort because IGRA testing requires blood draws and is often prohibitively expensive for large cohort studies, and TST is not readily available in South Africa. To validate the LTBI rate in our controls, we IGRA-tested (within 3 days of the participant's GeneXpert result), a random selection of our sample controls (n = 70) and found that they have an 87% LTBI rate (IGRA+). The highly significant model results, in combination with the IGRA+ subset of controls, suggest that our sampling strategy reliably categorizes TB disease risk.

We considered whether our sampling strategy displayed any indication of selection bias. We hypothesized that the risk of active TB should monotonically increase with age, reflecting cumulative lifetime exposure to the *M.tb* pathogen. Contrary to this, our findings revealed a non-monotonic relationship. Starting with the youngest age group (18–28 years), the empirical odds of active TB increased with age, peaking in the 29-38-year-old age group, followed by a progressive decline in empirical odds till the oldest age group (79–88 years) (Fig B in S1 Text). This unexpected pattern can be indicative of selection bias, possibly driven by survivor bias—where individuals dying from TB are absent from the study population at older age bins and/ or related to whether our controls are representative of the general population (sample-selection bias). Generally, selection bias is difficult to measure and mitigate [52–54], especially in

case-control studies where controls are recruited from a clinical setting. Recall bias in older adults could also lead to the lower observed empirical OR in older age bins; however, because TB treatment is a six-month course, and incomplete treatment regimens lead to relapse or life-threatening drug resistant TB, it is generally a life event that people remember with the exception of pediatric TB.

SES model effects

Age-specific TB risk varies across the lifespan. The greatest risk of TB is during infancy, decreasing through adolescence, then increasing and peaking between 25–35 years old followed by a decrease, and another peak after 65 years [48, 55]. Age was not a significant variable in the logistic regression except when interacting with SES. SES's protective effect on TB risk is most evident among 18–39-year-olds but the trend reverses and increases risk among the eldest individuals (>69 years; Fig 4A)—those who grew up and reached adulthood during Apartheid (Fig 4A). Higher SES increasing TB risk at older ages is contrary to findings in populations in the United States and Mexico [55]. This unique pattern may reflect South Africa's recent history of Apartheid and post-Apartheid societal and economic shifts. During Apartheid, individuals from historically marginalized backgrounds had limited career options, but some were able to become teachers, police officers, or nurses. Such occupations are associated with higher education requirements and would have facilitated access to larger salaries, transportation, and mobility; potentially leading to better diagnosis and treatment. Alternatively, the observed pattern of our interaction effect at older ages could be explained by selection bias or our operationalization of SES in this study. Highest completed level of education (e.g., grade, diploma, degree, etc.) is a blunt measure of SES, and does not fully capture all SES facets, including social, economic, and cultural capital [56, 57], and universal access to education increased post-Apartheid [58]. Additionally, we only sample from community health clinics, not private clinics, thereby missing a fraction of the SES spectrum.

Residence model effects

Consistent with previous research [21, 59–61], we find TB risk is associated with living in larger towns. In our prior work, mobility in the Northern and Western Cape populations changed over the past 3 generations, with the highest levels of mobility in the grandparental generation [62]. Therefore, we tested whether mobility (operationalized as a different birthplace and residence) affected TB risk. Individuals currently residing in towns (regardless of birthplace) had higher odds of active TB, compared to those born in towns that migrated to rural areas, and lifetime rural dwellers. Unexpectedly, the individuals with the lowest odds of active TB are those born in a town who move to a rural area (Fig 4B). When we returned results to the community, the clinic staff hypothesized that despite nationally standardized BCG vaccinations, rural areas may have lower vaccination rates (observation communicated by clinical staff in the study catchment), therefore those born in town benefit from a greater likelihood of greater BCG vaccine during childhood and low adult *M.tb* exposure living in rural areas. The benefits of the BCG vaccine, however, attenuate in adolescents and memory of childhood TB episodes suffer from recall bias. Another possibility is that the town-born and rural-residence group accrued more wealth in towns before moving to a rural area, affording a different lifestyle than their rural neighbors (e.g., afford larger homes, less crowding, cleaner cooking fuel, etc.). This unique combination of factors may explain why the town-born rural-residence group has even lower odds of active TB compared to the lifetime rural dwellers. Future work should consider collecting birthplace in addition to current residence to better identify TB risk as *M.tb* exposure varies across the lifespan.

Common risk factors

Across global studies, men are on average 1.7 times more likely to have TB [63–65]. Sex biases like this are common in other infectious diseases [66, 67] and are attributable to an intersection of sex (biological factors, e.g., immune function) and gender (social and behavioral factors, e.g., risk-taking behavior) [68]. Despite smoking not being a significant TB risk, we found 75.5% of men smoke compared to 55.8% of women, indicating at least some gender differences in risky behaviors in the Northern Cape population.

Smoking and alcohol consumption has been shown to increase TB risk and mortality in the Northern Cape and at the national level [18, 69–71]. In our models, smoking had the expected effect on TB risk and alcohol was excluded from our models due to high missingness. Self-reporting biases in observational studies like this one are a concern for variables like smoking, alcohol consumption, and SES measures [72]. Our sample, however, reports much higher levels of smoking compared to large-scale national surveys (e.g., [73], men: 75.5.% vs. 41%; women: 55.8% vs. 21%, respectively suggesting minimal self-report bias in our study. The weak effect of smoking observed from our models may be due to our method of binary classification. We collected fine-scale smoking phenotypes (Supplementary Methods in S1 text) but because of the high missingness of these phenotypes, we ultimately classified participants as Smokers/Non-smokers. This stratification may mask the heterogeneity of smoking behaviors such as casual and binge substance use or differences in the types of smoking materials consumed.

Ancestry & ethnicity

Finally, we highlight that our study included enrollment from 12 different clinics, some of which are more than 250 kilometers apart. We surveyed ethnicity and genetic ancestry to test for population structure in the sample. Such structure can confound analysis if genetic ancestry tracks differential host risk for progression to TB or if different ethnicities have different cultural norms. Previous studies have described the high proportion of Khoe-San ancestry in Northern Cape communities but these largely focused on descendant groups who identify as Khoe-San (e.g. the ≠Khomani San, the Nama, Karretjie) rather than the general population [74]. Here, we show the clinical study population to be admixed with 4 other distinct ancestries (Fig 3), demonstrative of recent historical events. These include the Bantu expansion into Southern Africa, European colonialism, the Dutch East India Company (aka VOC] slave trade, and the displacement and forced settlement of indigenous South African Khoe-San groups, especially in the last few generations in the Northern Cape [43, 75]. Although we do observe heterogeneity in ancestry across clinics, correcting for the top 10 genetic principal components did not change the logistic regression results (Fig D in S1 Text and S1 Data). To our knowledge, this is the first study to report ancestry proportions of clinical populations in the Northern Cape Province, South Africa. This work provides a baseline to design future studies, such as exploring host genetic correlates of active TB progression in this population (Supplementary Discussion in S1 text).

Conclusion

Active TB progression is a multifactorial process involving the environment, genetics, and their interaction [1, 7]. Our results from the NCTB cohort indicate that sociodemographic variables strongly impact active TB risk. Effects that are unique to the Northern Cape Province may reflect how changes in the pre- to post-apartheid environment modified social factors, such as SES and mobility, which in turn impacted lifetime TB risk.

Supporting information

S1 Checklist.
(DOCX)

S1 Data.
(XLSX)

S1 Text.
(DOCX)

Acknowledgments

Author Contributions

Conceptualization: Oshiomah P. Oyageshio, Justin W. Myrick.

Data curation: Oshiomah P. Oyageshio, Justin W. Myrick, Jamie Saayman, Lena van der Westhuizen, Dana R. Al-Hindi, Brenna M. Henn.

Formal analysis: Oshiomah P. Oyageshio, Noah Zaitlen, Brenna M. Henn.

Funding acquisition: Eileen G. Hoal, Marlo Möller, Brenna M. Henn.

Investigation: Oshiomah P. Oyageshio, Justin W. Myrick, Austin W. Reynolds, Noah Zaitlen, Caitlin Uren, Marlo Möller.

Methodology: Oshiomah P. Oyageshio, Justin W. Myrick, Brenna M. Henn.

Project administration: Justin W. Myrick, Jamie Saayman, Lena van der Westhuizen, Dana R. Al-Hindi, Austin W. Reynolds, Eileen G. Hoal, Caitlin Uren, Marlo Möller, Brenna M. Henn.

Resources: Lena van der Westhuizen, Austin W. Reynolds, Eileen G. Hoal, Marlo Möller, Brenna M. Henn.

Supervision: Justin W. Myrick, Jamie Saayman, Caitlin Uren, Marlo Möller, Brenna M. Henn.

Validation: Oshiomah P. Oyageshio, Justin W. Myrick.

Visualization: Oshiomah P. Oyageshio.

Writing – original draft: Oshiomah P. Oyageshio, Justin W. Myrick, Marlo Möller, Brenna M. Henn.

Writing – review & editing: Eileen G. Hoal, Caitlin Uren.

References

1. World Health Organization. Global Tuberculosis Report 2021 [Internet]. 2021 Oct [cited 2022 Aug 8]. Available from: https://www.who.int/publications/i/item/9789240037021

2. Sharma SK, Mohan A. Extrapulmonary Tuberculosis. Indian J Med Res. 2004 Nov 11; 120(4):316. PMID: 15520485

3. Stallings CL, Glickman MS. Is Mycobacterium tuberculosis stressed out? A critical assessment of the genetic evidence. Microbes Infect Inst Pasteur. 2010 Dec; 12(14–15):1091–1101.

4. Statistics South Africa. Mortality and Causes of death in South Africa: Findings from death notification 2018 [Internet]. 2018 [cited 2022 Apr 4]. Available from: https://www.statssa.gov.za/publications/P03093/P030932017.pdf

5. National Institute for Communicable Diseases. Microbiologically Confirmed Pulmonary TB—South Africa. 2019. TB Online Surveillance Dashboard. Available from: https://www.nicd.ac.za/tb-surveillance-dashboard/

6. Kim H, Tanser F, Tomita A, Vandormael A, Cuadros DF. Beyond HIV prevalence: identifying people living with HIV within underserved areas in South Africa. BMJ Glob Health. 2021 Apr 1; 6(4):e004089. https://doi.org/10.1136/bmjgh-2020-004089 PMID: 33883186

7. Glaziou P, Falzon D, Floyd K, Raviglione M. Global epidemiology of tuberculosis. Semin Respir Crit Care Med. 2013; 34(1):3–16. https://doi.org/10.1055/s-0032-1333467 PMID: 23460002

8. Lacerda SNB, De Abreu TRC, De Figueiredo TMRM, De Luna FDT, De Sousa MAN, De Abreu LC, et al. Individual and social vulnerabilities upon acquiring tuberculosis: A literature systematic review. Int Arch Med. 2014 Jul; 7(1):1–8.

9. Bloom BR, Atun R, Cohen T, Dye C, Fraser H, Gomez GB, et al. Tuberculosis. In: Holmes KK, Bertozzi S, Bloom BR, Jha P, editors. Major Infectious Diseases [Internet]. 3rd ed. Washington (DC): The International Bank for Reconstruction and Development / The World Bank; 2017 [cited 2024 Mar 18]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK525174/

10. Boelaert JR, Vandecasteele SJ, Appelberg R, Gordeuk VR. The effect of the host's iron status on tuberculosis. J Infect Dis. 2007 Jun 15; 195(12):1745–53. https://doi.org/10.1086/518040 PMID: 17492589

11. Chadha VK, Kumar P, Jagannatha PS, Vaidyanathan PS, Unnikrishnan KP. Average annual risk of tuberculous infection in India. Int J Tuberc Lung Dis. 2005 Jan; 9(1):116–8. PMID: 15675560

12. Hanifa Y, Grant AD, Lewis J, Corbett EL, Fielding K, Churchyard G. Prevalence of latent tuberculosis infection among gold miners in South Africa. Int J Tuberc Lung Dis. 2009 Jan; 13(1):39–46. PMID: 19105877

13. Isler MA, Rivest P, Mason J, Brassard P. Screening employees of services for homeless individuals in Montréal for tuberculosis infection. J Infect Public Health. 2013 Jun; 6(3):209–15.

14. Bruchfeld J, Correia-Neves M, Källenius G. Tuberculosis and HIV Coinfection. Cold Spring Harb Perspect Med. 2015 Jul; 5(7):a017871. https://doi.org/10.1101/cshperspect.a017871 PMID: 25722472

15. Bhargava A, Benedetti A, Oxlade O, Pai M, Menzies D. Undernutrition and the incidence of tuberculosis in India: national and subnational estimates of the population-attributable fraction related to undernutrition. Natl Med J India. 2014; 27(3):128–33. PMID: 25668081

16. Dhlakama H, Lougue S, Mwambi HG, Ogunsakin RE. A Multilevel Analysis of the Associated and Determining Factors of TB among Adults in South Africa: Results from National Income Dynamics Surveys 2008 to 2017. Int J Environ Res Public Health. 2022 Aug 25; 19(17):10611. https://doi.org/10.3390/ijerph191710611 PMID: 36078327

17. Dhlakama H, Lougue S. Bayesian Modelling of Tuberculosis Risk Factors in South Africa 2014. Int J Stat Med Res. 2017 Feb 27; 6(1):34–48.

18. Harling G, Ehrlich R, Myer L. The social epidemiology of tuberculosis in South Africa: A multilevel analysis. Soc Sci Med. 2008; 66(2):492–505. https://doi.org/10.1016/j.socscimed.2007.08.026 PMID: 17920743

19. Mahomed H, Hawkridge T, Verver S, Geiter L, Hatherill M, Abrahams DA, et al. Predictive factors for latent tuberculosis infection among adolescents in a high-burden area in South Africa. Int J Tuberc Lung Dis. 2011 Mar; 15(3):331–6. PMID: 21333099

20. Moyo S, Ismail F, Walt MV der, Ismail N, Mkhondo N, Dlamini S, et al. Prevalence of bacteriologically confirmed pulmonary tuberculosis in South Africa, 2017–19: a multistage, cluster-based, cross-sectional survey. Lancet Infect Dis. 2022 Aug 1; 22(8):1172–80. https://doi.org/10.1016/S1473-3099(22)00149-9 PMID: 35594897

21. Ncayiyana JR, Bassett J, West N, Westreich D, Musenge E, Emch M, et al. Prevalence of latent tuberculosis infection and predictive factors in an urban informal settlement in Johannesburg, South Africa: a cross-sectional study. BMC Infect Dis. 2016 Nov 8; 16(1):661. https://doi.org/10.1186/s12879-016-1989-x PMID: 27825307

22. Fok A, Numata Y, Schulzer M, FitzGerald MJ. Risk factors for clustering of tuberculosis cases: a systematic review of population-based molecular epidemiology studies. Int J Tuberc Lung Dis. 2008 May 1; 12(5):480–92. PMID: 18419882

23. WHO. Global Tuberculosis Report 2013. World Health Organization; 2013.

24. Diedrich CR, Flynn JL. HIV-1/mycobacterium tuberculosis coinfection immunology: how does HIV-1 exacerbate tuberculosis? Infect Immun. 2011 Apr; 79(4):1407–17. https://doi.org/10.1128/IAI.01126-10 PMID: 21245275

25. South African National Department of Health, South African Medical Research Council, Human Sciences Research Council, National Institute for Communicable Diseases, World Health Organization, United States Agency for International Development, et al. The First National TB Prevalence Survey | South Africa 2018 [Internet]. 2018 [cited 2021 Sep 11]. Available from: https://www.nicd.ac.za/wp-content/uploads/2021/02/TB-Prevalence-survey-report_A4_SA_TPS-Short_Feb-2021.pdf

26. Squires A. Epidemiology of tuberculosis in Northern Cape Province [Internet]. 2011. Available from: https://www.researchgate.net/profile/Adeline-Squires/publication/272821340_Epidemiology_of_tuberculosis_in_Northern_Cape_Province/links/54f004100cf2432ba657c798/Epidemiology-of-tuberculosis-in-Northern-Cape-Province.pdf

27. National Institute for Communicable Diseases [Internet]. Microbiologically Confirmed Pulmonary TB—Province. 2017. TB Online Surveillance Dashboard. Available from: https://www.nicd.ac.za/tb-surveillance-dashboard/

28. Verver S, Warren RM, Munch Z, Richardson M, et al. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. The Lancet. 2004 Jan 17; 363(9404):212–4.

29. 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. 2015 Jan 1; 211(1):53–61. https://doi.org/10.1093/infdis/jiu403 PMID: 25053739

30. Wood R, Liang H, Wu H, Middelkoop K, Oni T, Rangaka MX, et al. Changing prevalence of tuberculosis infection with increasing age in high-burden townships in South Africa. Int J Tuberc Lung Dis. 2010 Apr; 14(4):406–12. PMID: 20202297

31. Smith MH, Myrick JW, Oyageshio O, Uren C, Saayman J, Boolay S, et al. Epidemiological correlates of overweight and obesity in the Northern Cape Province, South Africa. PeerJ. 2023 Feb 9; 11:e14723. https://doi.org/10.7717/peerj.14723 PMID: 36788809

32. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. Am J Epidemiol. 2019 Dec 31; 188(12):2222–39. https://doi.org/10.1093/aje/kwz189 PMID: 31509183

33. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011 Dec 12; 45:1–67.

34. Rubin DB. Multiple Imputation for Nonresponse in Surveys [Internet]. John Wiley & Sons, Ltd; 1987 [cited 2024 Apr 19]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470316696.fmatter

35. Burgette LF, Reiter JP. Multiple Imputation for Missing Data via Sequential Regression Trees. Am J Epidemiol. 2010 Nov 1; 172(9):1070–6. https://doi.org/10.1093/aje/kwq260 PMID: 20841346

36. Heymans MW. psfmi: Prediction Model Pooling, Selection and Performance Evaluation Across Multiply Imputed Datasets [Internet]. Available from: https://mwheymans.github.io/psfmi/

37. Little RJ, Rubin DB. Statistical analysis with missing data. John Wiley & Sons; 2019 April 23.

38. Fox J. Effect Displays in R for Generalised Linear Models. J Stat Softw. 2003 Jul 22; 8:1–27.

39. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Sep; 19(9):1655. https://doi.org/10.1101/gr.094052.109 PMID: 19648217

40. Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinforma Oxf Engl. 2016 Sep 15; 32(18):2817–23. https://doi.org/10.1093/bioinformatics/btw327 PMID: 27283948

41. Statistics South Africa. Provincial profile: Northern Cape Community Survey 2016, Report 03-01-14 [Internet]. 2018. Available from: http://cs2016.statssa.gov.za/wp-content/uploads/2018/07/NorthernCape.pdf

42. Adhikari M. The Sons of Ham: Slavery and the Making of Coloured Identity. South Afr Hist J. 1992 Nov 1; 27(1):95–112.

43. Legassick M. Hidden Histories of Gordonia: Land dispossession and resistance in the Northern Cape, 1800–1990. NYU Press; 2016.

44. LaValley MP. Logistic Regression. Circulation. 2008 May 6; 117(18):2395–9. https://doi.org/10.1161/CIRCULATIONAHA.106.682658 PMID: 18458181

45. Baker PA. From Apartheid to Neoliberalism: Health Equity in Post-Apartheid South Africa. Int J Health Serv. 2010 Jan 1; 40(1):79–95. https://doi.org/10.2190/HS.40.1.e PMID: 20198805

46. Maphumulo WT, Bhengu BR. Challenges of quality improvement in the healthcare of South Africa post-apartheid: A critical review. Curationis. 2019 May 29; 42(1):e1–9. https://doi.org/10.4102/curationis.v42i1.1901 PMID: 31170800

47. Mhlanga D, Garidzirai R. The Influence of Racial Differences in the Demand for Healthcare in South Africa: A Case of Public Healthcare. Int J Environ Res Public Health. 2020 Jan; 17(14):5043. https://doi.org/10.3390/ijerph17145043 PMID: 32674256

48. Davies PDO. Risk factors for tuberculosis. Monaldi Arch Chest Dis. 2005 Mar; 63(1):37–46. https://doi.org/10.4081/monaldi.2005.656 PMID: 16035563

49. de Vries G, van Hest NAH, Baars HWM, Šebek MMGG, Richardus JH. Factors associated with the high tuberculosis case rate in an urban area. Int J Tuberc Lung Dis. 2010 Jul 1; 14(7):859–65. PMID: 20550769

50. Smith JP, Oeltmann JE, Hill AN, Tobias JL, Boyd R, Click ES, et al. Characterizing tuberculosis transmission dynamics in high-burden urban and rural settings. Sci Rep. 2022 Apr 26; 12(1):6780. https://doi.org/10.1038/s41598-022-10488-2 PMID: 35474076

51. Fofana AM, Moultrie H, Scott L, Jacobson KR, Shapiro AN, Dor G, et al. Cross-municipality migration and spread of tuberculosis in South Africa. Sci Rep. 2023 Feb 15; 13(1):2674. https://doi.org/10.1038/s41598-023-29804-5 PMID: 36792792

52. Szklo M, Nieto FJ. Epidemiology: Beyond the Basics. 3rd ed. Jones & Bartlett Learning; 2014.

53. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.

54. Woodward M. Epidemiology: Study Design and Data Analysis, 3rd ed. CRC Press; 2013.

55. Scordo JM, Aguillón-Durán GP, Ayala D, Quirino-Cerrillo AP, Rodríguez-Reyna E, Mora-Guzmán F, et al. A prospective cross-sectional study of tuberculosis in elderly Hispanics reveals that BCG vaccination at birth is protective whereas diabetes is not a risk factor. PLOS ONE. 2021 Jul; 16(7):e0255194–e0255194. https://doi.org/10.1371/journal.pone.0255194 PMID: 34324578

56. Khalatbari-Soltani S, Maccora J, Blyth FM, Joannès C, Kelly-Irving M. Measuring education in the context of health inequalities. Int J Epidemiol. 2022 Jun 1; 51(3):701–8. https://doi.org/10.1093/ije/dyac058 PMID: 35362521

57. Shavers VL. Measurement of socioeconomic status in health disparities research. J Natl Med Assoc. 2007 Sep; 99(9):1013–23. PMID: 17913111

58. Department of Education. South African Schools Act [Internet].1996 Nov [cited 2024 Mar 19]; Available from: https://www.gov.za/sites/default/files/gcis_document/201409/act84of1996.pdf

59. Beiranvand R, Karimi A, Delpisheh A, Sayehmiri K, Soleimani S, Ghalavandi S. Correlation Assessment of Climate and Geographic Distribution of Tuberculosis Using Geographical Information System (GIS). Iran J Public Health. 2016 Jan; 45(1):86–93. PMID: 27057526

60. Hoffner S, Hadadi M, Rajaei E, Farnia P, Ahmadi M, Jaberansari Z, et al. Geographic characterization of the tuberculosis epidemiology in iran using a geographical information system. Biomed Biotechnol Res J. 2018; 2(3):213.

61. Sikalengo G, Hella J, Mhimbira F, Rutaihwa LK, Bani F, Ndege R, et al. Distinct clinical characteristics and helminth co-infections in adult tuberculosis patients from urban compared to rural Tanzania. Infect Dis Poverty. 2018 Mar 24; 7(1):24. https://doi.org/10.1186/s40249-018-0404-9 PMID: 29580279

62. Reynolds A, Grote MN, Myrick JW, Al-Hindi DR, Siford RL, Mastoras M, et al. Persistence of matrilocal post-marital residence across multiple generations in Southern Africa. Human Nature. 2023 Jun; 34(2):295–323. https://doi.org/10.1007/s12110-023-09452-4 PMID: 37310564

63. Hertz D, Schneider B. Sex differences in tuberculosis. Semin Immunopathol. 2019 Mar; 41(2):225–37. https://doi.org/10.1007/s00281-018-0725-6 PMID: 30361803

64. Holmes CB, Hausler H, Nunn P. A review of sex differences in the epidemiology of tuberculosis. Int J Tuberc Lung Dis. 1998 Feb 1; 2(2):96–104. PMID: 9562118

65. Neyrolles O, Quintana-Murci L. Sexual Inequality in Tuberculosis. PLOS Med. 2009 Dec 22; 6(12):e1000199. https://doi.org/10.1371/journal.pmed.1000199 PMID: 20027210

66. World Health Organization(Western Pacific Region). Taking sex and gender into account in emerging infectious disease programme: an analytical framework [Internet]. 2007 [cited 2022 Aug 29]. Available from: https://www.who.int/publications-detail-redirect/9789290615323

67. Wizemann TM, Pardue ML. Exploring the Biological Contributions to Human Health [Internet]. National Academies Press (US); 2001 [cited 2022 Aug 29]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK222288/

68. Lawry LL, Lugo-Robles R, McIver V. Improvements to a framework for gender and emerging infectious diseases. Bull World Health Organ. 2021 Sep 1; 99(9):682–4. https://doi.org/10.2471/BLT.20.275636 PMID: 34475606

69. Peltzer K, Louw J, Mchunu G, Naidoo P, Matseke G, Tutshana B. Hazardous and Harmful Alcohol Use and Associated Factors in Tuberculosis Public Primary Care Patients in South Africa. Int J Environ Res Public Health. 2012 Sep; 9(9):3245–57. https://doi.org/10.3390/ijerph9093245 PMID: 23202681

70. Sitas F, Urban M, Bradshaw D, Kielkowski D, Bah S, Peto R. Tobacco attributable deaths in South Africa. Tob Control. 2004 Dec; 13(4):396–9. https://doi.org/10.1136/tc.2004.007682 PMID: 15564624

71. Wessels J, Walsh CM, Nel M. Smoking habits and alcohol use of patients with tuberculosis at Standerton Tuberculosis Specialised Hospital, Mpumalanga, South Africa. Health SA SA Gesondheid. 2019 Oct 8; 24:1146. https://doi.org/10.4102/hsag.v24i0.1146 PMID: 31934425

72. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc. 2016 May 4; 9:211–7. https://doi.org/10.2147/JMDH.S104807 PMID: 27217764

73. National Department of Health (NDoH), Statistics South Africa, South African Medical Research Coucil, ICF. South Africa Demographic and Health Survey 2016. [Internet]. [cited 2022 Aug 8]. Available from: https://dhsprogram.com/pubs/pdf/FR337/FR337.pdf

74. Uren C, Kim M, Martin AR, Bobo D, Gignoux CR, van Helden PD, et al. Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. Genetics. 2016 Sep 1; 204(1):303–14. https://doi.org/10.1534/genetics.116.187369 PMID: 27474727

75. Penn N. The Forgotten Frontier: Colonist and Khoisan on the Cape's Northern Frontier in the 18th Century. Juta and Company Ltd; 2005.