

**Title:** Strong Effect of Demographic Changes on Tuberculosis Susceptibility in South Africa

**Authors:** Oshiomah P. Oyageshio<sup>¶</sup>, Justin W. Myrick<sup>2¶</sup>, Jamie Saayman<sup>3</sup>, Lena van der Westhuizen<sup>3</sup>, Dana Al-Hindi<sup>4</sup>, Austin W. Reynolds<sup>5</sup>, Noah Zaitlen<sup>6</sup>, Caitlin Uren<sup>3,7</sup>, Marlo Möller<sup>3,7\*</sup>, Brenna M. Henn<sup>1,2,4\*</sup>

**Affiliations:**

- 1) Center for Population Biology, University of California, Davis, Davis, CA 95616, USA.
- 2) UC Davis Genome Center, University of California, Davis, Davis, CA 95616, USA.
- 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, CA 95616, USA.
- 5) Department of Anthropology, Baylor University, Waco, TX 76798, USA.
- 6) Department of Computational Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, 90095, USA.
- 7) Centre for Bioinformatics and Computational Biology, Stellenbosch University, Stellenbosch, South Africa.

¶ Equal authorship contribution

\* Corresponding authors: [bmhenn@ucdavis.edu](mailto:bmhenn@ucdavis.edu), [marlom@sun.ac.za](mailto:marlom@sun.ac.za)

Journal: PLoS Global Public Health

# Abstract

South Africa is among the world's top eight 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 tuberculosis (TB) in the Northern Cape Province, South Africa: an understudied TB endemic region with extreme TB incidence (645/100,000) and the lowest provincial population density. We leveraged the population's high TB incidence and community transmission to design a case-control study with population-based controls, reflecting 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 878 individuals (cases =374, controls =504) 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. Additionally, a subset of individuals were genotyped to determine genome-wide ancestry components. Male gender had the strongest effect on TB risk (OR: 2.87 [95% CI: 2.1-3.8]); smoking and alcohol consumption did not significantly increase TB risk. We identified two interactions: age by socioeconomic status (SES) and birthplace by residence locality on TB risk (OR = 3.05, p = 0.016) – where rural birthplace but town residence was the highest risk category. Finally, participants had a majority Khoe-San ancestry, typically greater than 50%. Epidemiological risk factors for this cohort differ from other global populations. The significant interaction effects reflect rapid changes in SES and mobility over recent generations and strongly impact TB risk in the Northern Cape of South Africa. Our models show that such risk factors combined explain 16% of the variance ( $r^2$ ) in case/control status.

## 46 Introduction

47 Tuberculosis (TB) is among the world's leading causes of death due to infectious disease,  
48 recently surpassed by COVID-19 (1). The TB causative agent, *Mycobacterium tuberculosis*  
49 (*M.tb*), is an obligate, exclusive *Homo sapiens* pathogen mainly infecting the lungs, and  
50 sometimes other organs (2,3). Determinants of active TB progression are multifaceted including  
51 human host genetics, nutrition, social and economic conditions, behavior, and sex-specific  
52 biology (1,4,5). The extent of these determinants' effects varies across and within populations,  
53 necessitating epidemiological studies in differing contexts and communities (5). These factors  
54 have also been shown to vary between low and high/intermediate-incidence populations, with  
55 lower odds ratios in high/intermediate-incidence populations (6). Here, we characterize the TB  
56 epidemiology of a district in the Northern Cape Province, South Africa, a TB-endemic region  
57 with relatively low HIV.

58 South Africa is amongst the top 30 'high burden' countries, burdened by TB, TB/HIV co-  
59 infection, and multi-drug resistance or rifampicin-resistant TB (MDR/RR-TB). TB is South  
60 Africa's leading natural cause of death (7) with an extremely high prevalence (852/100,000, (8))  
61 and accounts for 3.3% of all global TB cases (1). HIV is commonly identified as the leading risk  
62 factor for TB. In South Africa, 59% of TB patients on a TB programme (screened by a clinician  
63 and on TB medication) are co-infected with HIV. South Africa's first national TB prevalence  
64 survey (n=35,000), however, found only 28% of TB cases were co-infected with HIV(8). This  
65 discrepancy is partly explained by those with TB who go undetected, mainly symptomatic men  
66 not living with HIV who have limited clinical contact (8). Many individuals who are  
67 diagnostically TB+ may also go undetected because 78% of TB+ HIV- individuals exhibit only  
68 one or no classic TB symptoms (e.g. cough for two weeks, fever, night sweats, and weight loss;  
69 61% have no symptoms)(8).

70 In case-control studies, controls should have similar disease exposure profiles to the  
71 cases. Population-based controls risk differential disease exposure, a concern in low-incidence  
72 populations that can bias statistical associations. However, in high-incidence populations with  
73 well-characterized disease burdens and transmission, population controls greatly improve  
74 statistical power (9). In South Africa, TB is community spread moreover than household (10,11)  
75 and TB latency increases with age, reaching 80% by age 30 (12–15), an epidemiological scenario  
76 that ensures cases and controls have approximate disease exposure profiles.

77 *Mycobacterium tuberculosis* has a long coevolutionary history with different human  
78 populations likely leading to population-specific genetic signatures (2,16). TB susceptibility  
79 phenotypes have a heritability of 11-92% (17), yet few critical genetic variants have replicated  
80 across genome-wide association studies (GWAS) (18), potentially reflecting these population-  
81 specific signatures. This result has spurred several studies to examine the relationship between  
82 genetic ancestry and TB risk (19–23). For instance, native Amerindian ancestry was shown to be  
83 a risk factor for TB progression in an admixed Amazonian population, and genetic variants in  
84 Peruvian populations have been associated with early active TB progression (19–21). A major  
85 challenge in identifying genetic risk factors associated with TB progression is decoupling the  
86 social and environmental effects that accompany ancestry. To this end, Asgari et al. controlled  
87 for environmental effects (e.g., sanitation, water supply, and socioeconomic status [SES]) and  
88 found indigenous Peruvian ancestry to be an independent, significant predictor of TB  
89 progression. Chimusa et al. also corrected for SES and demonstrated an association between  
90 Khoe-San ancestry and TB progression in South Africa (22).

91 In this study, we investigated ancestry proportions as well as several common TB  
92 epidemiological variables identified in earlier studies (24). Smoking, alcohol consumption, and  
93 intravenous drug use have independently been associated with TB. Meta-analyses have found  
94 alcohol use and smoking (25,26), and specifically heavy alcohol use (26–28) to increase TB risk,  
95 though not always consistently. Our study is part of the Northern Cape Tuberculosis Project  
96 (NCTB), investigating the human-host genetics of TB among admixed Khoe-San descent  
97 populations in rural or peri-urban communities. Characterizing the TB epidemiology in this  
98 region will identify nongenetic risk factors that can serve as control variables in future genetic  
99 studies of TB risk.

## 101 **Methods**

### 103 **Research ethics statement**

104 This study has been approved by the Health Research Ethics Committee (HREC) of  
105 Stellenbosch University (N11/07/210A) and the Northern Cape Department of Health  
106 (NC2015/008). All participants were adults (18 years and older) and provided written informed  
107 formal consent. Authors Justin W. Myrick, Jamie Saayman, Lena van der Westhuizen and Marlo  
108 Möller had access to identifiable information about participants as they were directly involved in  
109 data collection or database management. Access to these records commenced on 26th January  
110 2016, and is still ongoing as it is an integral part of the ongoing Northern Cape Tuberculosis  
111 Project (NCTB).

### 113 **Study Design and Recruitment**

114 Participants (18 years and older) provided written informed consent and were recruited  
115 from 12 community health clinics from the ZF Mgcawu district in the Northern Cape Province of  
116 South Africa from 26th January 2016 - 15 May 2017, and 11 December 2018 - 11 March 2020.  
117 Community health clinics are the front line for TB screening and treatment, visited by 87% of  
118 people who seek TB care (8). TB nurses referred patients with suspected TB (with  $\geq 2$  TB  
119 symptoms: cough for  $\geq 2$  weeks, night sweats, weight loss, and fever  $\geq 2$  weeks or a TB contact)  
120 and TB patients to our on-site RAs. All study participants took a clinic-administered sputum  
121 GeneXpert Ultra test for active TB at the time of the study interview and provided saliva for  
122 genotyping. Clinic medical charts were accessed by a staff research nurse to record GeneXpert  
123 test results and verify HIV status and TB history.

### 125 **Case-Control Assignment**

126 Cases and controls were assigned reckoning the participant's medical charts and self-  
127 reported data (see Fig. 1). Cases include anyone with active pulmonary TB in their lifetime *and*  
128 are HIV-negative and followed two tracks: 1) Clinically confirmed active TB (n= 343) and 2)  
129 self-reported past TB episode(s) (n=228). GeneXpert results, diagnostic test date, TB strain (drug  
130 resistance), and TB medication regimen were used to determine clinically confirmed progression  
131 to active TB. Past TB episodes are self-reported, mainly due to older medical charts not reliably  
132 available, discarded, or difficult to locate by clinic staff.

133 Controls are patients with suspected TB who have a negative GeneXpert Ultra result and  
134 have no history of active pulmonary TB at the time of study enrollment and are largely assumed  
135 to be latently infected with *M.tb* (LTBI). A majority of the population in high TB burden South

136 African suburbs are LTBI, 88% by ages 31-35 (12,13) and studies have consistently shown LTBI  
137 in South Africa to be above 75% by age 25, increasing across adulthood (14). Our population-  
138 control design relies on population-wide TB exposure, as traditional screening methods,  
139 tuberculin skin test (TST) and interferon-gamma release assay (IGRA; e.g., QuantiFERON), are  
140 limited both in the concordance and positive predictive value (29,30). IGRA and TST are used to  
141 infer *M.tb* infection, but cannot be used to determine previous exposure to the bacterium. Certain  
142 individuals living in high *M.tb* exposed populations test persistently negative for these tests and  
143 do not develop active disease, but display *Mtb*-specific antibody titres. These individuals are  
144 known as “resisters” or “early clearers” (31,32).

145 Our exclusion criteria removed participants with unknown TB or HIV status, as well as  
146 individuals with dual HIV and TB infections.

147  
148 **Fig 1. Case-Control Decision Tree.** Study participants were  
149 categorized as cases or controls based on medical record  
150 information and self-reported data. All participants were  
151 GeneXpert tested for active TB infection at the time of enrollment.  
152 Past TB episodes were self-reported and cross-referenced with  
153 medical records when available.

154

## 155 **Study covariates**

156 We collected demographic information that included date of birth, place of birth, current  
157 residence, self-identified gender, self-reported ethnic identity, and parental ethnic identities.  
158 Behavioral variables include smoking and alcohol consumption (See Supplementary Materials in  
159 S1 Text). In our analyses, we only used binary measures for smoking and alcohol (“Do you  
160 smoke?”, “Do you drink alcohol?”). Residence and birthplace locations are categorized as rural  
161 ( $\leq 2000$  people) and town ( $> 2,000$  people). Population size was derived from the South African  
162 census and when census data was absent, e.g., a farm, we used Google Earth (earth.google.com)  
163 to estimate population size based on the number of dwellings. Age was used as a continuous  
164 variable for all analyses and binned for calculating empirical odds (see Fig 2B). SES was  
165 operationalized as someone’s number of years of education, i.e., the highest completed level of  
166 education. McKenzie et al. have shown education level, in this dataset, positively predicts body  
167 mass index in TB controls, tracking access to resources and food security (33).

168

169 **Fig 2. Case-Control status shifts across Age groups.** A)  
170 Overlapping density plots of age distribution stratified by TB  
171 status (n= 878). At the oldest and youngest ages, most of our study  
172 participants are cases whilst at middle-age groups, the majority  
173 are controls. B) Empirical odds of active TB by age group. The x-  
174 axis bins our participants into 7 age groups and the y-axis: the  
175 empirical odds of active TB. Empirical odds are calculated by  
176 dividing the number of controls divided by the number of cases in  
177 each age bin. The size of the dots corresponds to the sample size  
178 of the age group. Our data reveal a signal of survivor bias. Since  
179 age is a cumulative measure of exposure, the empirical odds of TB  
180 should increase with age. This pattern is observed from our  
181 youngest age group up to age 58. The empirical odds of TB

182 progressively decrease after age 58. Older age groups are biased  
183 towards controls due to the mortality of TB.

184

## 185 **Data Analyses**

186 Statistical analyses were performed in R (version 4.0.2). We calculated Pearson  
187 correlations with the R package *ggcorrplot*. All categorical variables were numerically coded to  
188 “0” and “1”. Classification models for our binary, qualitative dependent variable (“case”/  
189 “control”) included logistic regression and random forest—a machine learning classifier robust  
190 to non-linear associations and unknown variable interactions (34) (see Supplementary Materials  
191 in S1 Text). Random forest is a growing analytic tool in epidemiology (35–37). The coefficients  
192 of the logistic regression models were converted to odds ratios using the R package *gtools* (38),  
193 and marginal effects were plotted using the R package *effects* (39). Each model was Bonferroni  
194 corrected by dividing, 0.05, by the number of variables in said model.

195

196 Our first model, the common risk factor model (n=878), includes seven covariates known  
197 to be common risk factors for TB.

198

199 TB Status ~ gender + smoking + drinking + diabetes + residence + age + SES

200

201 Health disparities are one of the many consequences of apartheid in South Africa (40,41).  
202 The end of apartheid improved social mobility and educational access, however, health  
203 disparities in the Northern Cape still remain (42). To capture the effect of lived experience vis-à-  
204 vis Apartheid on TB outcomes we designed the “SES model” (n=878). This model includes the  
205 common risk factor model and interacts with age and SES. Age is kept as a continuous variable  
206 because Apartheid was not a historically binary event.

207

208 TB Status ~ common risk factor model + age \* SES

209

210 Residing in an urban or rural environment is an established risk factor for TB status. In  
211 the “residence model”, we test the relationship between current residence and birthplace  
212 residence on TB status. Here, we build on the common risk factor model to include an interaction  
213 between current residence and birthplace. Setting this interaction allows us to examine four  
214 patterns, namely: rural birthplace to urban residence, urban birthplace to rural residence, lifetime  
215 rural residence, and lifetime urban residence.

216

217 TB Status ~ common risk factor model + residence \* birthplace

218

## 219 **Genetic Data Processing & Ancestry Estimation**

220 Genetic data processing involved DNA extraction from saliva samples, genotyping for >2  
221 million SNPs, common variant calling with GenomeStudio, rare variant calling with zCall, and  
222 further data cleaning using plink2 with specific parameters (Supplementary methods in S1  
223 Text). Prior to genetic ancestry estimation, SNPs out of Hardy-Weinberg equilibrium (--hwe 0.001)  
224 and rare alleles (--maf 0.01) were removed from the dataset. The dataset was also pruned for  
225 linkage disequilibrium (--indep-pairwise 200 25 0.4). Individuals from Luhya, Maasai, Himba,  
226 British, Palestinian, Chinese, Bangladeshi, Tamil, Ju|’hoansi San, Khomani San, Nama  
227 populations were used as reference groups. Global ancestry estimates were calculated using

228 ADMIXTURE v1.13(43). This was done in groups of maximally unrelated individuals to avoid  
229 biasing the ancestry estimates. ADMIXTURE was run for  $k=5$  on unsupervised mode for each of  
230 the running groups. After matching clusters, we merged ancestry estimates across all running  
231 groups, averaging individuals that appeared in multiple running groups using pong (44)

232

## 233 **Results**

234

### 235 **TB case-control classification**

236 1,126 participants were partitioned into preliminary cases, preliminary controls, and  
237 unverified TB status (571,504, and 51 respectively; Table C in S1 Text). After excluding,  
238 participants with unverified TB status, preliminary cases with unverified HIV status, and  
239 participants co-infected with TB and HIV, 878 participants remained in the study (374 cases and  
240 504 controls; Table A in S1 Text).

241

242

### 243 **Socio-behavioral covariates and demographics**

244 Men and women were equally represented in the dataset (422:441, respectively, Table A  
245 in S1 Text). Men were more likely to drink alcohol ( $r = -0.14$ ,  $p < 0.05$ ; Fig I in S1 Text) and  
246 smoke ( $r = -0.22$ ,  $p < 0.05$ ; Fig I in S1 Text). Most of our participants smoked (66%) and 45%  
247 drank alcohol; smoking and drinking were moderately correlated with each other ( $r = 0.36$ ,  $p <$   
248  $0.05$ ; Fig I in S1 Text). Women were more likely to have diabetes ( $r = 0.12$ ,  $p < 0.05$ ; Fig I in S1  
249 Text) and, on average, had more education than men (female mean= 8.3 years, male mean = 7.7  
250 years).

251 Cases and controls had similar distributions for age (mean = 43.1, SD =13.2 and mean  
252 =42.4, SD =15.2, respectively, Table A in S1 Text). “Age” is defined here as the age at the time  
253 of study enrollment and importantly, is a cumulative outcome: that is, it includes cases who  
254 currently and/or previously had TB, not the age of the TB episode. Age also captures the amount  
255 of time someone is exposed to TB. The empirical odds of active TB in our data reveal a signature  
256 of survivorship bias (Fig. 2B). We use the number of years of education as a proxy for “SES”.  
257 The mean educational attainment is 8 years, equivalent to completing primary school, and  
258 similar between rural areas and towns (ANOVA,  $p > 0.1$ ). In the ZF Mgcawu District census  
259 (45) 13% of people have not completed primary school compared to 25.3% of our participants.  
260 Age was moderately correlated with SES ( $r = -0.5$ ,  $p < 0.05$ ; Fig I in S1 Text) such that older  
261 participants tended to have lower SES.

262

### 263 **Ethnicity and Khoe-San Ancestry**

264 Genetic ancestry analyses were performed for 159 participants (see Supplementary  
265 Methods in S1 Text) from the Northern Cape Tuberculosis Project on host-genetic susceptibility  
266 to TB. To our knowledge, this is the first study to report ancestry proportions of a clinical  
267 population in the Northern Cape Province, South Africa. Khoe-San ancestry varied across clinic  
268 locations (Fig. 3A) but remained the majority ancestry at each site (mean = 56%), followed by  
269 Bantu-speaking African ancestry (mean = 21%), European ancestry (mean = 16%), South Asian  
270 ancestry (mean = 5%), and East Asian ancestry (mean = 2%) (Fig. 3B).

271

272 **Fig 3. Khoe-San Ancestry is the Primary Genetic Ancestry in Clinics**  
 273 **from the Northern Cape, South Africa.** A subset of participants  
 274 (n=159) was genotyped for preliminary ancestry analysis. A) The  
 275 study population is admixed with 5 distinct ancestries with the  
 276 Southern African indigenous Khoe-San ancestry being the largest  
 277 proportion of ancestry across all study sites. (B) Although Khoe-  
 278 San ancestry is the largest proportion of ancestry in our sample,  
 279 it varies significantly across study sites.

280  
 281 Individuals were asked to self-identify their ethnicity without prompting. 88.4% of  
 282 participants (both TB cases and controls) self-identify as, coloured, followed by 4.2% as a Khoe-  
 283 San ethnicity (e.g., Nama, San), 4.6 % as Tswana, 1.3 % as Xhosa, and 1.9 % as “other”. Whilst  
 284 we acknowledge that in some contexts the term, coloured, has derogatory connotations, it is a  
 285 recognized ethnicity and used culturally in South Africa. People who self-identify using this term  
 286 have different ancestries of different geographic origins, including the indigenous Khoe-San  
 287 groups (e.g., Khoekhoe, San), Bantu-speaking, European, Indian, Malaysian (Southeast Asian)  
 288 slaves, or people of mixed ancestry and their descendants (46).  
 289

### 290 **Logistic Regression Results**

291 We designed three logistic regression models (47) to examine the risk factors’ odds ratios  
 292 for the binary dependent variable, TB case/control. The common risk factor model included age,  
 293 SES, gender, residence, smoking, diabetes, and alcohol as covariates. The SES model extended  
 294 the common risk factor model to include an interaction between age and SES. Lastly, the  
 295 residence model extended the common risk factor model to include an interaction between  
 296 birthplace and current residence. The SES model (AIC = 1099; pseudo  $r^2$  = 17%, Table 1)  
 297 performed slightly better than the common risk factor model (AIC= 1108; pseudo  $r^2$  =16%, Table  
 298 1). The residence model had a similar pseudo  $r^2$  (15% (Table B in S1 Text)) as the other two,  
 299 however, we could not compare their AICs due to different sample sizes. All significance levels  
 300 were Bonferroni corrected. This was carried out by dividing 0.05 by the number of variables  
 301 used in the model.  
 302

303 **Table 1:** Odds ratios and p-values for the Demographic and Socio  
 304 Behavioral Variables used in the Common Risk Factor Model and SES  
 305 Model  
 306  
 307

	Common Risk Factor Model		SES Model	
	Odds Ratio (CI)	p-value	Odds Ratio (CI)	p-value
<b>Intercept</b>	0.305 [0.13, 0.72]	0.007	2.237 [0.46, 11.18]	0.32
<b>Gender -</b>	2.87 [2.10,	7.842e-12	2.85 [2.12,	6.545e-



<b>Male</b>	3.80]		3.85]	12
<b>Years of Education (SES)</b>	0.95 [0.90, 1.00]	0.03	0.75 [0.63, 0.88]	0.0007
<b>Age</b>	0.996 [0.99, 1.00]	0.55	0.959 [0.931, 0.986]	0.003
<b>Drinks Alcohol - Yes</b>	1.05 [0.77, 1.43]	0.77	1.02 [0.75, 1.40]	0.88
<b>Diabetes - Yes</b>	1.36 [0.73, 2.50]	0.35	1.37 [0.74, 2.52]	0.31
<b>Current Residence - Rural (reference)</b>	1		1	
<b>Current Residence - Town</b>	2.88 [2.07, 4.02]	5.316e-10	2.91 [2.09, 4.09]	3.912e-10
Smoker - Yes	1.31 [0.94, 1.84]	0.114	1.27 [0.90, 1.78]	0.171
Years of Education * Age			1.005 [ 1.002, 1.009]	0.003
<b>N</b>	<b>878</b>			
<b>Significant p-value</b>	<b>p &lt; 0.007</b>		<b>p &lt; 0.006</b>	
<b>AIC</b>	<b>1108</b>		<b>1099</b>	
<b>Pseudo-R<sup>2</sup> (Cragg-Uhler)</b>	<b>0.15</b>		<b>0.17</b>	

308

### 309 **Gender, Alcohol, Smoking, and Diabetes**

310 Males have three times the odds of active TB than females (OR = 2.85, p < 0.001; Table  
311 1 and Fig. 4). All logistic regression models showed insufficient statistical evidence for smoking  
312 (common risk factor model: OR = 1.31, p = 0.11; Table 1, Table B in S1 Text), alcohol

313 consumption (common risk factor model: OR = 1.05, p = 0.77; Table 1 and Table B in S1 Text)  
314 and diabetes (common risk factor model: OR =1.36, p=0.32; Table 1 and Table B in S1 Text) on  
315 TB risk. Despite the lack of significance, we note that smoking had an effect size in the expected  
316 direction (Fig. 4).

317  
318 **Figure 4. Effect Plots demonstrating the relationship between**  
319 **Active TB Status and A) Gender, B) Current Residence and C)**  
320 **Smoking.** These plots are reported from the best-performing  
321 logistic regression model (SES model). Y-axes for all panels  
322 show the odds of active TB. We find that the odds of active TB  
323 are 3 times higher in Males. Individuals currently residing in  
324 Towns have about 2.5 times higher odds of active TB as compared  
325 to individuals currently residing in rural areas. Smoking  
326 slightly increases the odds of active TB but is not  
327 statistically significant.

### 328 329 **Age Interacts with SES**

330 In the common risk factor model, age (OR= 0.996 [0.99, 1.00], p=0.55) and SES (OR =  
331 0.947, p=0.0324; see Table 1) have no effect on TB risk. To examine this unexpected finding, we  
332 interacted age with years of education (proxy for SES). SES significantly affects TB status  
333 depending on age group (OR =1.005, p = 0.004, Table 1). The effect takes on a U-shaped  
334 relationship across ages, such that higher SES at younger ages (18-39 years old) is protective  
335 against TB, and higher SES at older ages (>59 years) increases risk (Fig. 5). Middle-aged  
336 individuals (40-59 years old) show no relationship between age and SES on TB risk (Fig. 5).

337  
338 **Fig 5. Logistic regression interaction plots.** A) The odds of  
339 active TB by education level vary across age groups (shown above  
340 by the different color lines). More years of education decreases  
341 the odds of active TB in younger age groups, but this pattern  
342 reverses in the oldest age groups. In middle-aged individuals,  
343 there is no relationship between age and years of education. B)  
344 Effect plot from the residence model visualizing an interaction  
345 term between birthplace residence and current residence.  
346 Regardless of birthplace, the odds of active TB is highest in  
347 individuals who currently reside in towns. Individuals born in  
348 towns and currently residing in rural areas have the lowest odds  
349 of active TB.

### 350 351 **TB Risk is Highest in Towns**

352 The odds of active TB were significantly higher for people residing in towns (common  
353 risk factor model: OR = 2.88 [2.07-4.03], p <0.0001; Table 1 and Fig. 4). For the residence  
354 model, we analyzed the impact of moving between rural areas and towns during an individual's  
355 lifetime (birthplace by residence) on TB status. We expected to see a difference in odds for TB  
356 risk between life-long residents and those who have moved between locales. Under such a  
357 model, lifelong rural dwellers would have the lowest odds and lifelong town dwellers would  
358 have the highest odds. We set an interaction term between current residence and birthplace

359 classified into town/rural; this interaction was marginally significant (OR = 3.05, p = 0.016;  
360 Table B in S1 Text). Our results show that regardless of birthplace, current residence in a town  
361 area increases the risk of active TB (Fig. 3B). Interestingly, individuals who were born in a town  
362 and later moved to rural areas are even more protected than individuals born and currently  
363 residing in rural areas (Fig. 3B).  
364

## 365 **Random Forest Modeling**

366 Similar to logistic regression, random forest is a binary classifier yet differs in that is  
367 robust against non-linear associations and unknown interactions (34). Random forest utilizes a  
368 permutation-based approach to generate a hierarchical list of important variables but is unable to  
369 quantify the “significance” between an independent and dependent variable.

370 5000 subsets of our dataset were used to grow 5000 classification trees using baseline  
371 variables as predictors for active TB status. The model assigned gender, current residence, and  
372 SES respectively as the overall top important independent variables (Fig II-A in S1 Text). Age,  
373 diabetes, alcohol, and smoking were classified as uninformative predictors for TB. The random  
374 forest model also stratified the variable importance by cases and controls. Gender was the top  
375 predictor for case status, followed by current residence and SES (Fig II-B in S1 Text). SES was  
376 the top predictor for control status followed by gender and current residence (Fig II-C in S1  
377 Text). Interestingly, age had some predictive relevance for case status but was the worst-  
378 performing predictor for controls (Fig II-C in S1 Text). The model had an overall “out-of-bag”  
379 misclassification rate of 38%, with misclassification lower in controls (controls =30%, cases  
380 48%; Supplementary Materials in S1 Text).  
381

## 382 **Discussion**

383  
384 This present work represents the largest TB epidemiological study on a Northern Cape  
385 clinical population (n=878). In this study, we demonstrate the utility of population-based  
386 controls when disease exposure is known and transmission is community-spread (48) as seen in  
387 other studies in low-resource, high-burdened countries (9,49). Logistic regression and random  
388 forest models both show gender and residence as significant and important TB risk predictors.  
389 Random forest assigned SES as an important variable, and SES was only significant when  
390 interacting with age in logistic regressions. Neither smoking, alcohol consumption, nor diabetes  
391 is associated with increased TB risk in any model. Two logistic regression models, interacting  
392 SES by age (SES model), and birthplace by residence (residence model), had similar explanatory  
393 power, improving on the common risk factor model. This study demonstrates a possible unique  
394 historical context to South Africa, (post-)Apartheid differential effects between  
395 sociodemographic and health outcomes.

396 Age and TB risk have a general inverted U-shape relationship. During childhood, infants  
397 are at the greatest risk of TB decreasing through adolescence, increasing between 25-35 years  
398 old followed by a decrease, and another peak after 65 years (50,51). In our study population  
399 ( $\geq 18$  years), the empirical odds of active TB increase with age and peak in the 49-58-year-old  
400 age group, followed by a steady decline in empirical odds after age 58 until the oldest age group  
401 (Fig. 2B). This drop in empirical odds after age 58 is most likely due to the mortality of  
402 individuals with TB, potentially a signal of survivor bias (52). This interpretation is seen in the  
403 shifting proportions of cases and controls across age groups (Fig. 2A). From ages 21 - 58, most  
404 of the population are cases, and from ages 59 - 88 most of the population controls (Fig. 2A).

405 Age was neither a significant (logistic regression) nor an important (random forest)  
406 variable except when interacted with SES. SES's protective effect on TB risk is most prominent  
407 among 18-39 year-olds and becomes a risk factor among the eldest individuals (>65 years; Fig.  
408 5A)—those who grew up and reached adulthood during Apartheid (Fig. 5A). Higher SES  
409 increasing TB risk is contrary to findings in populations in the United States and Mexico (51).  
410 This unique pattern may reflect South Africa's recent history of Apartheid and post-Apartheid  
411 societal and economic shifts. During Apartheid, individuals from historically marginalized  
412 backgrounds had limited career options, but some were able to become teachers, police officers,  
413 or nurses. Such occupations are associated with higher education requirements and would have  
414 facilitated access to larger salaries, transportation, and mobility.

415 Higher SES could result in apparent greater odds of TB because these individuals would  
416 have had better access to healthcare both facilitating diagnosis and treatment. Universal access to  
417 education increased post-Apartheid but given the wide variation of years of education among the  
418 youngest generations, it likely still covaries with SES. Given the unusual interaction here  
419 between age and years of education, future work should validate additional SES measures to  
420 resolve mechanisms of TB risk.

421 Consistent with previous research (53–56), we find TB risk is associated with living in  
422 larger towns. In our prior work, mobility in the Northern and Western Cape populations changed  
423 over the past 3 generations with the highest levels of mobility in the grandparental generation  
424 (57). Therefore, we tested whether mobility (different birthplace and residence) affected TB risk.  
425 As expected, the protective effect of living rurally vanishes when someone moves to a larger  
426 town. Further, the individuals with the lowest TB risk are those born in a town and move to a  
427 rural area. These findings are consistent with TB exposure nearing ~ 90% by 25-30 years old  
428 (13), with transmission occurring via community contacts during adolescence and adulthood. We  
429 hypothesize that those born in towns who later moved to a rural area benefit from both BCG  
430 vaccination and decreased adult exposure thereby overall decreasing their odds of TB. BCG  
431 vaccination is standard for children in South Africa, however, children in rural areas may have  
432 lower vaccination rates (observation communicated by clinical staff in the study catchment).  
433 Future work should consider collecting birthplace in addition to current residence to better  
434 identify TB risk.

435 Invariably across studies, men are on average 1.7 times more likely to have TB (58–60).  
436 Sex biases like this are common in other infectious diseases (61,62) and are attributable to an  
437 intersection of sex (biological factors, e.g., immune function) and gender (social and behavioral  
438 factors, e.g., risk-taking behavior) (63). Despite smoking not being a significant TB risk, we  
439 found 75.5% of men smoke compared to 55.8% of women, indicating at least some gender  
440 differences in risky behaviors in the Northern Cape population.

441 Smoking and alcohol consumption has been shown to increase TB risk and mortality in  
442 the Northern Cape and at the national level (64–67). In our models smoking had the expected  
443 effect on TB risk and alcohol consumption had no effect. Both variables lacked statistical power  
444 in regression models and failed to meet any level of importance in the random forest model. Self-  
445 reporting biases in observational studies like this one are a concern for variables like smoking,  
446 alcohol consumption, and SES measures (68). Our sample, however, reports much higher levels  
447 of smoking compared to large-scale national surveys (e.g., (69), men: 75.5% vs. 41%; women:  
448 55.8% vs. 21%, respectively suggesting minimal self-report bias in our study. It is possible that  
449 these weak effects of smoking and alcohol observed from our models are due to our method of  
450 binary classification. We collected fine-scale smoking and alcohol phenotypes (Supplementary

451 Methods in S1 Text) but because of the high missingness of these phenotypes, we ultimately  
452 classified participants as Smokers/Non-smokers and Drinkers/Non-Drinkers. This stratification  
453 may mask the heterogeneity of drinking and smoking behaviors such as casual and binge  
454 substance use or differences in the types of alcohol and smoking materials consumed. Further TB  
455 epidemiological studies in the Northern Cape should explore these smoking and alcohol  
456 phenotypes in more detail.  
457 Active TB progression is a multifactorial process involving the environment, genetics, and their  
458 interaction (1,4). Our results from the NCTB cohort indicate that sociodemographic variables  
459 strongly impact active TB risk. Effects that are unique to the Northern Cape Province may reflect  
460 how changes in the pre- to post-apartheid environment modified social factors, such as SES and  
461 mobility, which in turn impacted lifetime TB risk. This work provides a baseline to design well-  
462 informed future studies, such as exploring host genetic correlates of active TB progression in this  
463 population (Supplementary Discussion in S1 Text).

464

## 465 **Declarations of Funding**

466 This work was funded by NIH grant R35GM133531 to BMH. The content is solely the  
467 responsibility of the authors and does not necessarily represent the official views of the National  
468 Institutes of Health. The South African government also partially funded this research through the  
469 South African Medical Research Council and the National Research Foundation.

470

## 471 **Conflicts of Interest**

472 All authors declare no conflict of interest.

473

## 474 **Acknowledgments**

475 We would like to thank all the participant communities in the Northern Cape for their  
476 continued trust and support in helping us undertake this project. We would also like to thank our  
477 community research assistants and translators who assisted in data collection for the project. We  
478 are grateful to Prof Faadiel Essop, Dr. Desiree Petersen, Prof Eileen Hoal, and Prof Leslie Swartz  
479 for a close reading of this manuscript. Finally, we want to thank the Department of Health in the  
480 Northern Cape Province, South Africa for their continued support of the project.

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495  
496  
497  
498  
499  
500  
  
501  
  
502  
  
503  
  
504  
  
505  
506  
507  
508  
509  
  
510  
511  
  
512  
513  
514  
515  
  
516  
517  
  
518  
519  
520  
  
521  
522  
523  
  
524  
525  
526  
  
527  
528

## References

1. WHO. Global Tuberculosis Report 2021 [Internet]. 2021 Oct [cited 2022 Aug 8]. Available from: <https://www.who.int/publications/i/item/9789240037021>
2. Brites D, Gagneux S. Co-evolution of Mycobacterium tuberculosis and Homo sapiens. *Immunol Rev.* 2015 Mar;264(1):6–24.
3. Sharma SK, Mohan A. Extrapulmonary Tuberculosis. *Indian J Med Res* [Internet]. 2004 [cited 2022 Nov 11];120(4)(316). Available from: <https://www.proquest.com/openview/330577dc52a107765d6adb9b1168c6e6/1?pq-origsite=gscholar&cbl=37533>
4. Glaziou P, Falzon D, Floyd K, Raviglione M. Global epidemiology of tuberculosis. *Semin Respir Crit Care Med.* 2013;34(1):3–16.
5. Lacerda SNB, De Abreu Temoteo RC, 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.
6. Fok A, Numata Y, Schulzer M, FitzGerald MJ. Risk factors for clustering of tuberculosis cases: a systematic review of population-based molecular epidemiology studies [Review Article]. *Int J Tuberc Lung Dis.* 2008 May 1;12(5):480–92.
7. 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>
8. South African National Department of Health, South African Medical Research Council, Human Sciences Research Council, National Institute for Communicable Diseases, World

- 529 Health Organization, United States Agency for International Development, et al. The First  
530 National TB Prevalence Survey | South Africa 2018 [Internet]. 2018 [cited 2021 Sep 11].  
531 Available from:  
532 [https://www.google.com/search?client=safari&rls=en&q=the+first+national+tb+prevalence+](https://www.google.com/search?client=safari&rls=en&q=the+first+national+tb+prevalence+survey+south+africa+2018&ie=UTF-8&oe=UTF-8)  
533 [survey+south+africa+2018&ie=UTF-8&oe=UTF-8](https://www.google.com/search?client=safari&rls=en&q=the+first+national+tb+prevalence+survey+south+africa+2018&ie=UTF-8&oe=UTF-8)
- 534 9. Duchen D, Vergara C, Thio CL, Kundu P, Chatterjee N, Thomas DL, et al. Pathogen  
535 exposure misclassification can bias association signals in GWAS of infectious diseases when  
536 using population-based common control subjects. *Am J Hum Genet.* 2023 Feb;110(2):336–  
537 48.
- 538 10. Verver S, Warren RM, Munch Z, Richardson M, et al. Proportion of tuberculosis  
539 transmission that takes place in households in a high-incidence area. *The Lancet.* 2004 Jan  
540 17;363(9404):212–4.
- 541 11. Middelkoop K, Mathema B, Myer L, Shashkina E, Whitelaw A, Kaplan G, et al.  
542 Transmission of Tuberculosis in a South African Community With a High Prevalence of  
543 HIV Infection. *J Infect Dis.* 2015 Jan 1;211(1):53–61.
- 544 12. Gallant CJ, Cobat A, Simkin L, Black GF, Stanley K, Hughes J, et al. Impact of age and sex  
545 on mycobacterial immunity in an area of high tuberculosis incidence. *Int J Tuberc Lung Dis.*  
546 2010 Aug;14(8):952–9.
- 547 13. Bunyasi EW, Schmidt BM, Abdullahi LH, Mulenga H, Tameris M, Luabeya A, et al.  
548 Prevalence of latent TB infection and TB disease among adolescents in high TB burden  
549 countries in Africa: a systematic review protocol. *BMJ Open.* 2017 Mar 1;7(3):e014609.
- 550 14. Wood R, Liang H, Wu H, Middelkoop K, Oni T, Rangaka MX, et al. Changing prevalence of  
551 tuberculosis infection with increasing age in high-burden townships in South Africa. *Int J*  
552 *Tuberc Lung Dis Off J Int Union Tuberc Lung Dis.* 2010 Apr;14(4):406–12.
- 553 15. Uys P, Brand H, Warren R, Spuy G van der, Hoal EG, Helden PD van. The Risk of  
554 Tuberculosis Reinfection Soon after Cure of a First Disease Episode Is Extremely High in a  
555 Hyperendemic Community. *PLOS ONE.* 2015 Dec 9;10(12):e0144487.
- 556 16. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human  
557 populations. *Nat Rev Genet.* 2014 Apr;15(6):379–93.
- 558 17. Möller M, Kinnear CJ, Orlova M, Kroon EE, van Helden PD, Schurr E, et al. Genetic  
559 Resistance to Mycobacterium tuberculosis Infection and Disease. *Front Immunol* [Internet].  
560 2018 [cited 2022 Nov 11];9. Available from:  
561 <https://www.frontiersin.org/articles/10.3389/fimmu.2018.02219>
- 562 18. Uren C, Hoal EG, Möller M. Mycobacterium tuberculosis complex and human coadaptation:  
563 a two-way street complicating host susceptibility to TB. *Hum Mol Genet.* 2021 Apr  
564 26;30(R1):R146–53.

- 565 19. Leal DF da VB, Silva MNS da, Fernandes DCR de O, Rodrigues JCG, Barros MC da C,  
566 Pinto PD do C, et al. Amerindian genetic ancestry as a risk factor for tuberculosis in an  
567 amazonian population. *PLOS ONE*. 2020 Jul 16;15(7):e0236033.
- 568 20. Luo Y, Suliman S, Asgari S, Amariuta T, Baglaenko Y, Martínez-Bonet M, et al. Early  
569 progression to active tuberculosis is a highly heritable trait driven by 3q23 in Peruvians. *Nat*  
570 *Commun*. 2019;10(1):1–10.
- 571 21. Asgari S, Luo Y, Huang CC, Zhang Z, Calderon R, Jimenez J, et al. Higher native Peruvian  
572 genetic ancestry proportion is associated with tuberculosis progression risk. *Cell Genomics*.  
573 2022 Jul 13;2(7):100151.
- 574 22. Chimusa ER, Zaitlen N, Daya M, Möller M, Helden PD van, Nicola JM, et al. Genome-wide  
575 association study of ancestry-specific TB risk in the South African coloured population.  
576 *Hum Mol Genet*. 2014;23(3):796–809.
- 577 23. Daya M, van der Merwe L, van Helden PD, Möller M, Hoal EG. The role of ancestry in TB  
578 susceptibility of an admixed South African population. *Tuberculosis*. 2014 Jul 1;94(4):413–  
579 20.
- 580 24. Nava-Aguilera E, Andersson N, Harris E, Mitchell S, Hamel C, Shea B, et al. Risk factors  
581 associated with recent transmission of tuberculosis: systematic review and meta-analysis. *Int*  
582 *J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis*. 2009 Jan;13(1):17–26.
- 583 25. Bates MN, Khalakdina A, Pai M, Chang L, Lessa F, Smith KR. Risk of tuberculosis from  
584 exposure to tobacco smoke: A systematic review and meta-analysis. *Arch Intern Med*. 2007  
585 Feb;167(4):335–42.
- 586 26. Rehm J, Samokhvalov AV, Neuman MG, Room R, Parry C, Lönnroth K, et al. The  
587 association between alcohol use, alcohol use disorders and tuberculosis (TB). A systematic  
588 review. *BMC Public Health*. 2009 Dec;9(1):450.
- 589 27. Fiske CT, Hamilton CD, Stout JE. Alcohol use and clinical manifestations of tuberculosis. *J*  
590 *Infect*. 2009 May;58(5):395–401.
- 591 28. Imtiaz S, Shield KD, Roerecke M, Samokhvalov AV, Lönnroth K, Rehm J. Alcohol  
592 consumption as a risk factor for tuberculosis: meta-analyses and burden of disease. *Eur*  
593 *Respir J*. 2017 Jul;50(1):1700216.
- 594 29. Barnes PF. Weighing Gold or Counting Spots. *Am J Respir Crit Care Med*. 2006  
595 Oct;174(7):731–2.
- 596 30. Stout JE, Menzies D. Predicting Tuberculosis Does the IGRA Tell the Tale? *Am J Respir*  
597 *Crit Care Med*. 2008 May 15;177(10):1055–7.
- 598 31. Kroon EE, Kinnear CJ, Orlova M, Fischinger S, Shin S, Boolay S, et al. An observational  
599 study identifying highly tuberculosis-exposed, HIV-1-positive but persistently TB, tuberculin



- 600 and IGRA negative persons with *M. tuberculosis* specific antibodies in Cape Town, South  
601 Africa. *EBioMedicine*. 2020 Nov 1;61:103053.
- 602 32. Verrall AJ, Netea MG, Alisjahbana B, Hill PC, van Crevel R. Early clearance of  
603 *Mycobacterium tuberculosis*: a new frontier in prevention. *Immunology*. 2014  
604 Apr;141(4):506–13.
- 605 33. Smith MH, Myrick JW, Oyageshio O, Uren C, Saayman J, Boolay S, et al. Epidemiological  
606 correlates of overweight and obesity in the Northern Cape Province, South Africa. *PeerJ*.  
607 2023 Feb 9;11:e14723.
- 608 34. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the  
609 Epidemiologist. *Am J Epidemiol*. 2019 Oct 21;kwz189.
- 610 35. Iwendi C, Bashir AK, Peshkar A, Sujatha R, Chatterjee JM, Pasupuleti S, et al. COVID-19  
611 Patient Health Prediction Using Boosted Random Forest Algorithm. *Front Public Health*  
612 [Internet]. 2020 [cited 2022 Aug 29];8. Available from:  
613 <https://www.frontiersin.org/articles/10.3389/fpubh.2020.00357>
- 614 36. Loef B, Wong A, Janssen NAH, Strak M, Hoekstra J, Picavet HSJ, et al. Using random  
615 forest to identify longitudinal predictors of health in a 30-year cohort study. *Sci Rep*. 2022  
616 Jun 20;12(1):10372.
- 617 37. Ooka T, Johno H, Nakamoto K, Yoda Y, Yokomichi H, Yamagata Z. Random forest  
618 approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale  
619 health check-up data in Japan. *BMJ Nutr Prev Health* [Internet]. 2021 Jun 1 [cited 2022 Aug  
620 29];4(1). Available from: <https://nutrition.bmj.com/content/4/1/140>
- 621 38. Bolker B, Warnes G, Lumley T. *gtools: Various R Programming Tools*. R package version  
622 3.9.3 [Internet]. 2022. Available from: <https://CRAN.R-project.org/package=gtools>
- 623 39. Fox J. Effect Displays in R for Generalised Linear Models. *J Stat Softw*. 2003 Jul 22;8:1–27.
- 624 40. Baker PA. From Apartheid to Neoliberalism: Health Equity in Post-Apartheid South Africa.  
625 *Int J Health Serv*. 2010 Jan 1;40(1):79–95.
- 626 41. Maphumulo WT, Bhengu BR. Challenges of quality improvement in the healthcare of South  
627 Africa post-apartheid: A critical review. *Curationis*. 2019 May 29;42(1):e1–9.
- 628 42. Mhlanga D, Garidzirai R. The Influence of Racial Differences in the Demand for Healthcare  
629 in South Africa: A Case of Public Healthcare. *Int J Environ Res Public Health*. 2020  
630 Jan;17(14):5043.
- 631 43. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated  
632 individuals. *Genome Res*. 2009 Sep;19(9):1655.

- 633 44. Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. t-SNE: fast analysis and  
634 visualization of latent clusters in population genetic data. *Bioinforma Oxf Engl*. 2016 Sep  
635 15;32(18):2817–23.
- 636 45. Statistics South Africa. Provincial profile: Northern Cape Community Survey 2016, Report  
637 03-01-14. Private Bag X44, Pretoria, 0001; 2018.
- 638 46. Adhikari M. The Sons of Ham: Slavery and the Making of Coloured Identity. *South Afr Hist*  
639 *J*. 1992 Nov 1;27(1):95–112.
- 640 47. LaValley MP. Logistic Regression. *Circulation*. 2008 May 6;117(18):2395–9.
- 641 48. Munch Z, Van Lill SWP, Booysen CN, Zietsman HL, Enarson DA, Beyers N. Tuberculosis  
642 transmission patterns in a high-incidence area: a spatial analysis. *Int J Tuberc Lung Dis Off J*  
643 *Int Union Tuberc Lung Dis*. 2003 Mar;7(3):271–7.
- 644 49. Lienhardt C, Fielding K, Sillah J, Bah B, Gustafson P, Warndorff D, et al. Investigation of  
645 the risk factors for tuberculosis: a case–control study in three countries in West Africa. *Int J*  
646 *Epidemiol*. 2005 Aug 1;34(4):914–23.
- 647 50. Davies PDO. Risk factors for tuberculosis. *Monaldi Arch Chest Dis*. 2005 Mar;63(1):37–46.
- 648 51. Scordo JM, Aguillón-Durán GP, Ayala D, Quirino- Cerrillo AP, Rodríguez-Reyna E, Mora-  
649 Guzmán F, et al. A prospective cross-sectional study of tuberculosis in elderly Hispanics  
650 reveals that BCG vaccination at birth is protective whereas diabetes is not a risk factor.  
651 *PLOS ONE*. 2021 Jul;16(7):e0255194–e0255194.
- 652 52. Swanson DM, Anderson CD, Betensky RA. Hypothesis Tests for Neyman’s Bias in Case-  
653 Control Studies. *J Appl Stat*. 2018;45(11):1956–77.
- 654 53. Beiranvand R, Karimi A, Delpisheh A, Sayehmiri K, Soleimani S, Ghalavandi S. Correlation  
655 Assessment of Climate and Geographic Distribution of Tuberculosis Using Geographical  
656 Information System (GIS). *Iran J Public Health*. 2016 Jan;45(1):86–93.
- 657 54. Hoffner S, Hadadi M, Rajaei E, Farnia P, Ahmadi M, Jaberansari Z, et al. Geographic  
658 characterization of the tuberculosis epidemiology in iran using a geographical information  
659 system. *Biomed Biotechnol Res J BBRJ*. 2018;2(3):213.
- 660 55. Ncayiyana JR, Bassett J, West N, Westreich D, Musenge E, Emch M, et al. Prevalence of  
661 latent tuberculosis infection and predictive factors in an urban informal settlement in  
662 Johannesburg, South Africa: a cross-sectional study. *BMC Infect Dis*. 2016 Nov 8;16(1):661.
- 663 56. Sikalengo G, Hella J, Mhimbira F, Rutaihwa LK, Bani F, Ndege R, et al. Distinct clinical  
664 characteristics and helminth co-infections in adult tuberculosis patients from urban compared  
665 to rural Tanzania. *Infect Dis Poverty*. 2018 Mar 24;7(1):24.
- 666 57. Reynolds A, Grote MN, Myrick JW, Al-Hindi DR, Siford RL, Mastoras M, et al. Persistence  
667 of matrilineal post-marital residence across multiple generations in Southern Africa

- 668 [Internet]. SocArXiv; 2022 [cited 2022 Nov 11]. Available from:  
669 <https://osf.io/preprints/socarxiv/7qfns/>
- 670 58. Hertz D, Schneider B. Sex differences in tuberculosis. *Semin Immunopathol*. 2019  
671 Mar;41(2):225–37.
- 672 59. Holmes CB, Hausler H, Nunn P. A review of sex differences in the epidemiology of  
673 tuberculosis. *Int J Tuberc Lung Dis*. 1998 Feb 1;2(2):96–104.
- 674 60. Neyrolles O, Quintana-Murci L. Sexual Inequality in Tuberculosis. *PLOS Med*. 2009 Dec  
675 22;6(12):e1000199.
- 676 61. WHO (Western Pacific Region). Taking sex and gender into account in emerging infectious  
677 disease programme : an analytical framework [Internet]. 2007 [cited 2022 Aug 29].  
678 Available from: <https://www.who.int/publications-detail-redirect/9789290615323>
- 679 62. Wizemann TM, Pardue ML. Exploring the Biological Contributions to Human Health  
680 [Internet]. National Academies Press (US); 2001 [cited 2022 Aug 29]. Available from:  
681 <https://www.ncbi.nlm.nih.gov/books/NBK222288/>
- 682 63. Lawry LL, Lugo-Robles R, McIver V. Improvements to a framework for gender and  
683 emerging infectious diseases. *Bull World Health Organ*. 2021 Sep 1;99(9):682–4.
- 684 64. Harling G, Ehrlich R, Myer L. The social epidemiology of tuberculosis in South Africa: A  
685 multilevel analysis. *Soc Sci Med*. 2008;66(2):492–505.
- 686 65. Peltzer K, Louw J, Mchunu G, Naidoo P, Matseke G, Tutshana B. Hazardous and Harmful  
687 Alcohol Use and Associated Factors in Tuberculosis Public Primary Care Patients in South  
688 Africa. *Int J Environ Res Public Health*. 2012 Sep;9(9):3245–57.
- 689 66. Sitas F, Urban M, Bradshaw D, Kielkowski D, Bah S, Peto R. Tobacco attributable deaths in  
690 South Africa. *Tob Control*. 2004 Dec;13(4):396–9.
- 691 67. Wessels J, Walsh CM, Nel M. Smoking habits and alcohol use of patients with tuberculosis  
692 at Standerton Tuberculosis Specialised Hospital, Mpumalanga, South Africa. *Health SA SA*  
693 *Gesondheid*. 2019 Oct 8;24:1146.
- 694 68. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment  
695 methods. *J Multidiscip Healthc*. 2016 May 4;9:211–7.
- 696 69. National Department of Health (NDoH), Statistics South Africa, South African Medical  
697 Research Council, ICF. South Africa Demographic and Health Survey 2016. [Internet]. [cited  
698 2022 Aug 8]. Available from: <https://dhsprogram.com/pubs/pdf/FR337/FR337.pdf>

699  
700  
701  
702

703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717

## S1 Text

719 Supplementary Methods

720 Supplementary Results

721 Supplementary Discussion

722 **Table A.** Descriptive Statistics of study variables, stratified by  
723 case/control status

724 **Table B.** Odds ratios and p-values for the Demographic and Socio  
725 Behavioral Variables used in the Residence Model

726 **Table C.** Descriptive statistics of Raw Dataset (Before imputation  
727 and implementing HIV exclusion criteria

728 **Table D:** Missingness and Imputation Metrics

729

730

731 **Fig I. Correlations Among Demographic and Medical Variables for**  
732 **Entire Cohort.** Pearson correlation coefficients were calculated  
733 for select demographic, behavioral, and medical covariates in our  
734 dataset of 878 individuals. Correlations with significant p-values  
735 ( $p < 0.05$ ) are denoted with an asterisk.

736

737 **Fig II. Hierarchical list of important variables from Random**  
738 **Forest Model in A) All individuals (n= 878) B) Cases (n =374) and**  
739 **C) Controls (n =504).** Random subsets of all 7 variables on the y-  
740 axis were used to grow 5000 trees to classify participants into  
741 cases and controls. Mean decrease accuracy was computed by the  
742 differences in classification error between the "out-of-bag"  
743 dataset and a randomly permuted sample (*Supplementary methods in*  
744 *S1 text*). Variables with higher mean decrease accuracy are most  
745 important for case-control classification. Predictor variables  
746 with mean decrease accuracy values close to zero have no effect on  
747 classification while negative values worsen the ability of the  
748 model to classify TB status.



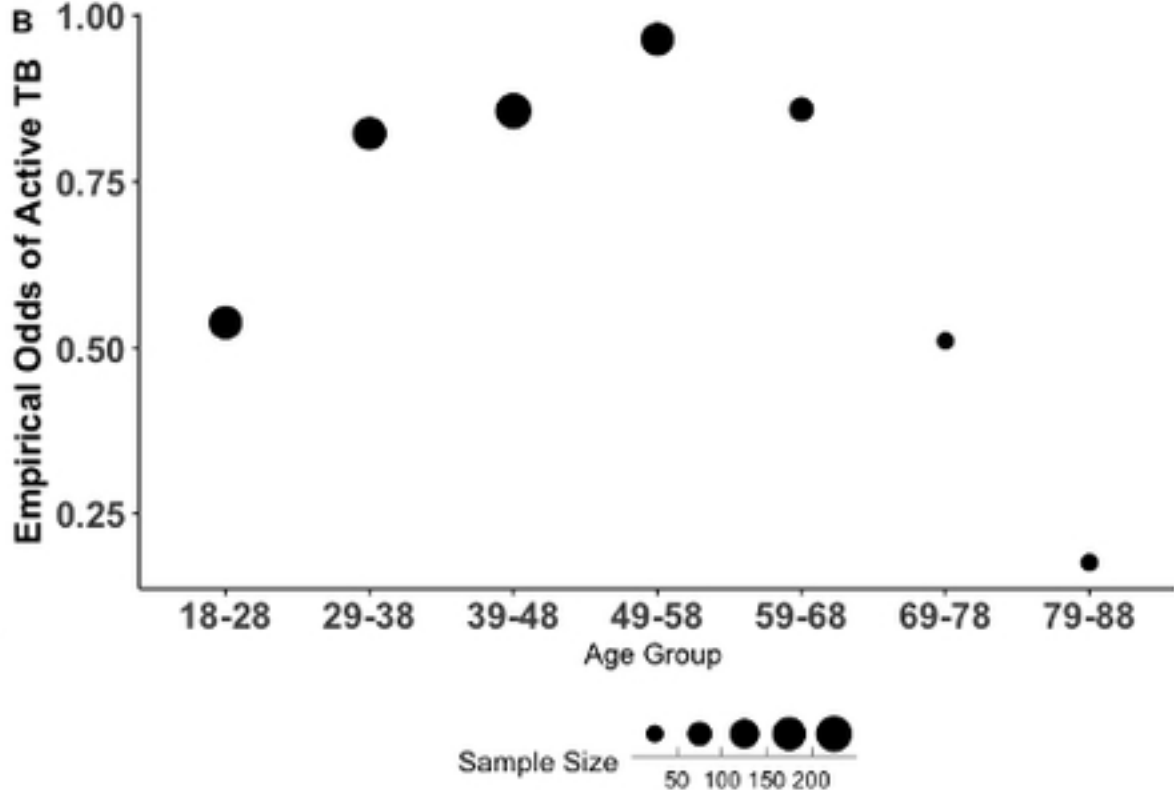
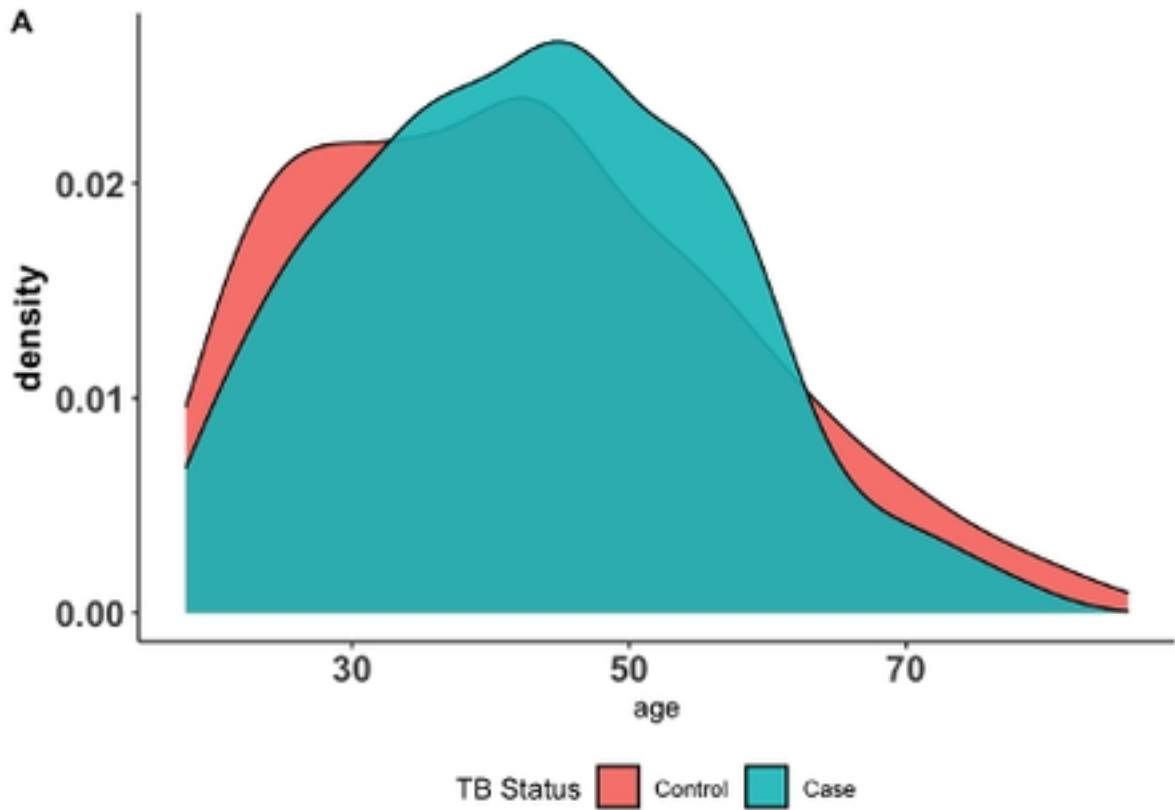


Figure 2

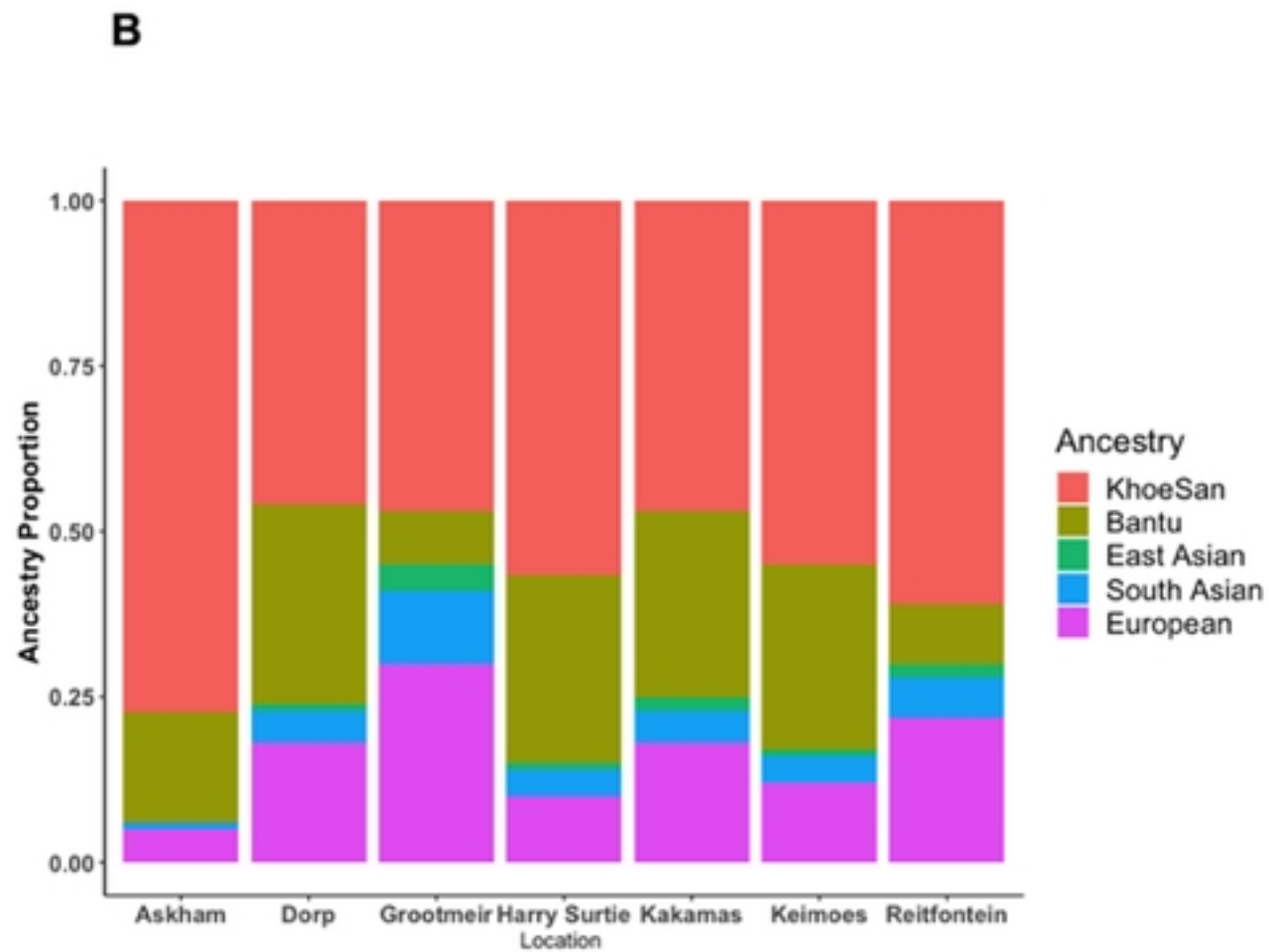
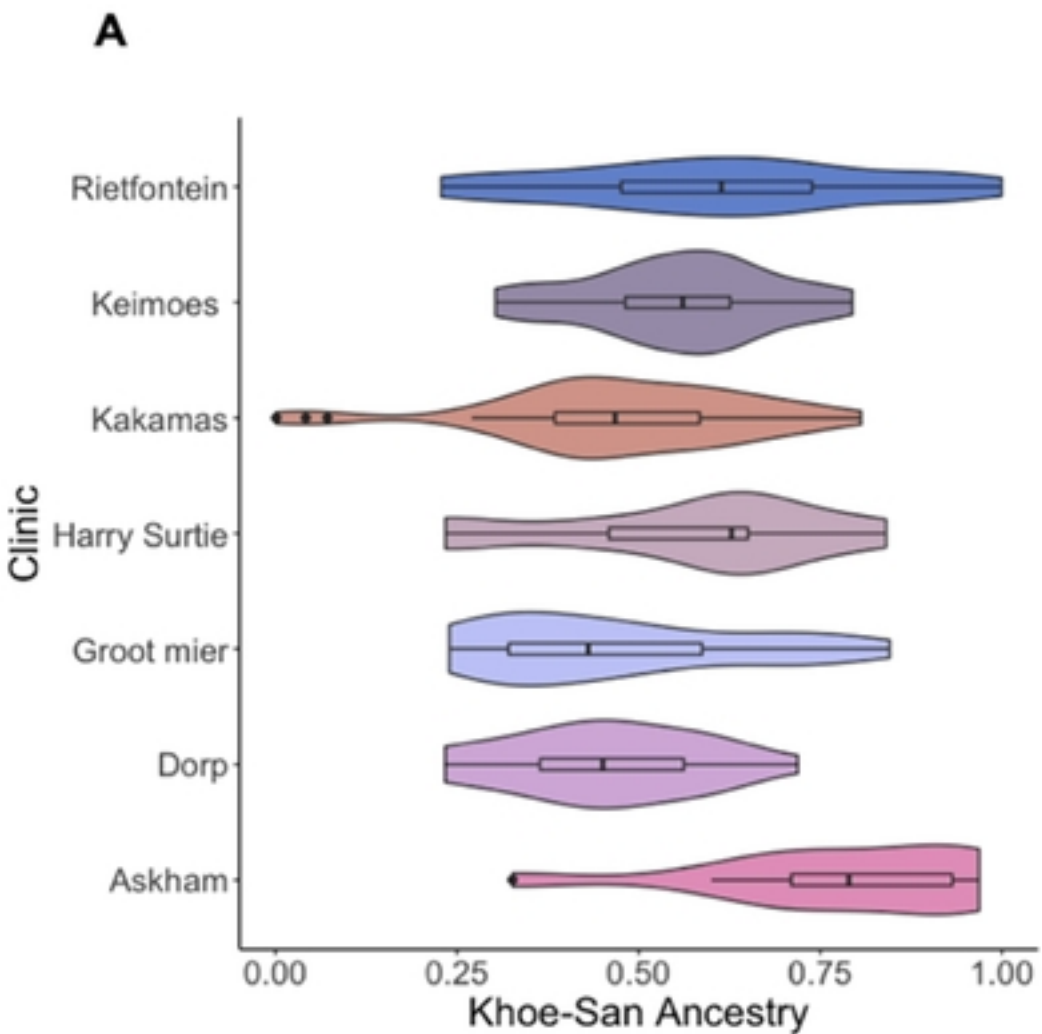


Figure 3

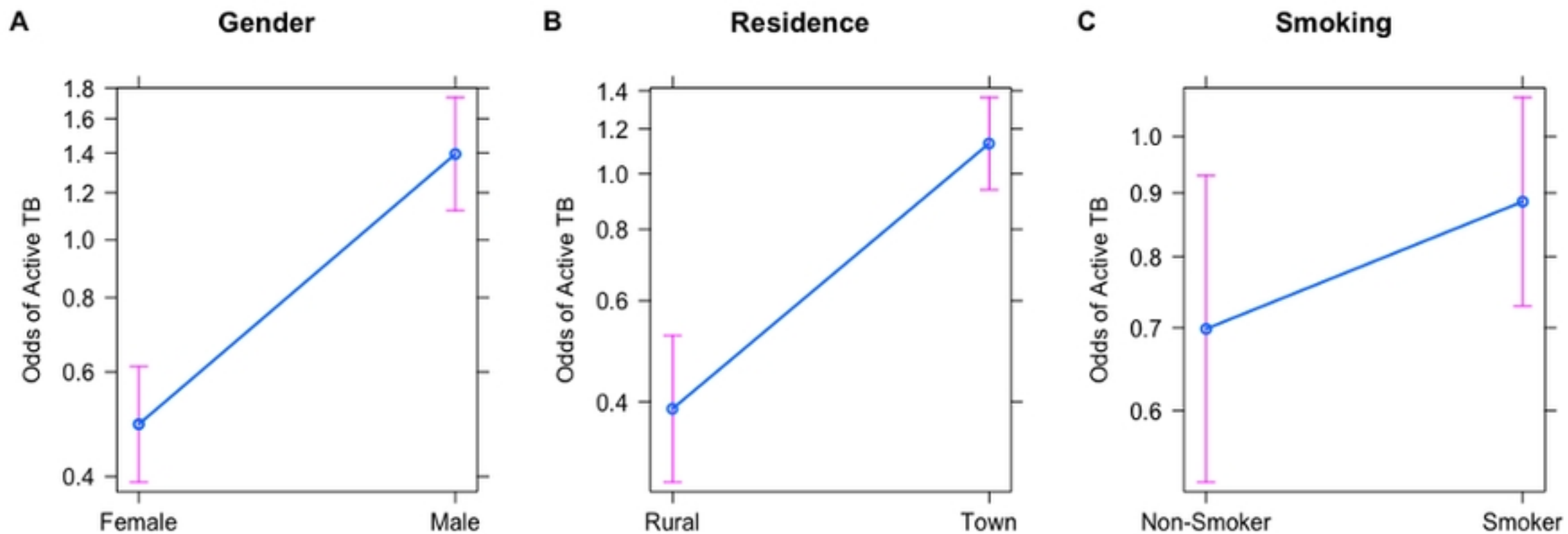


Figure 4



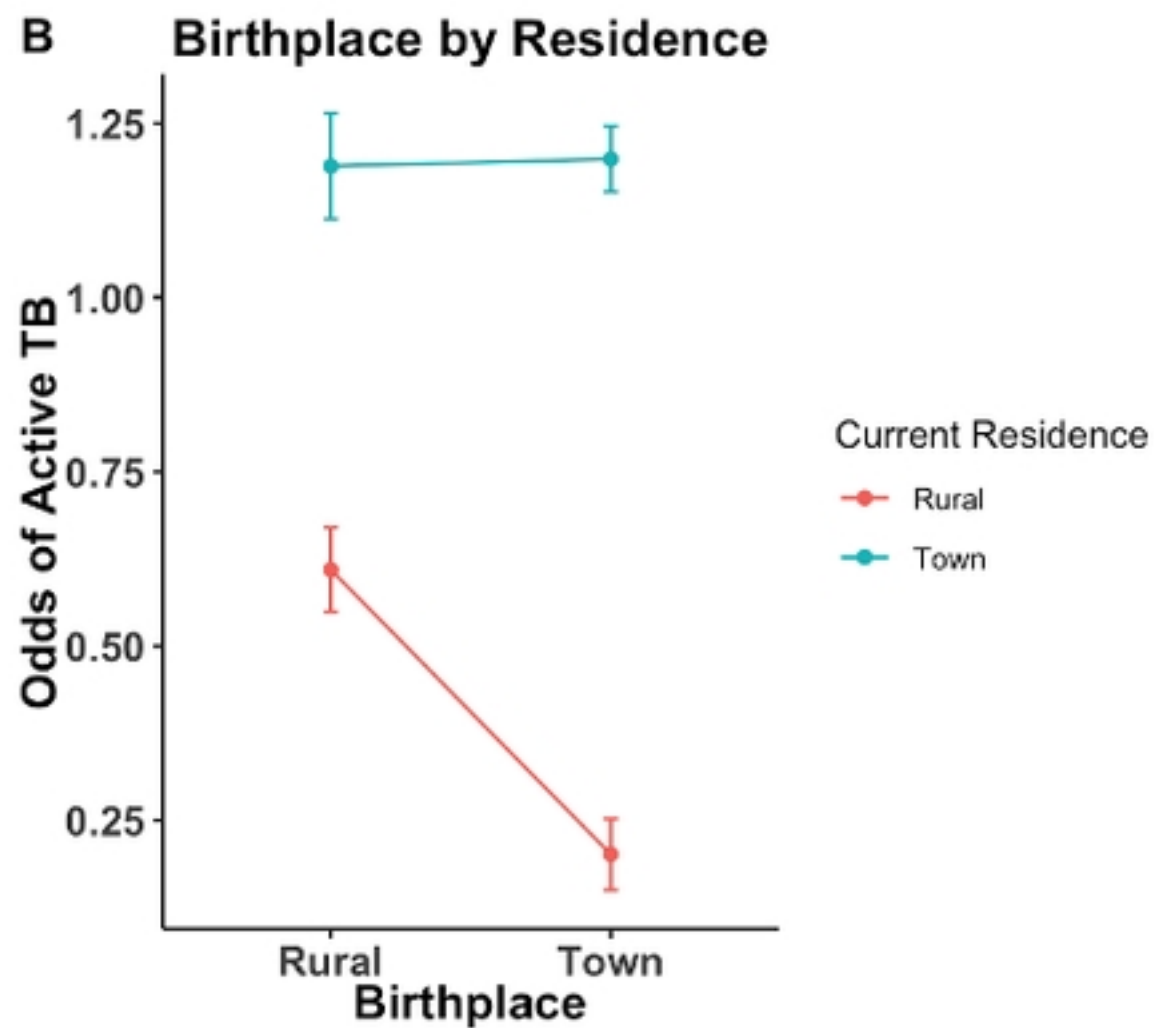
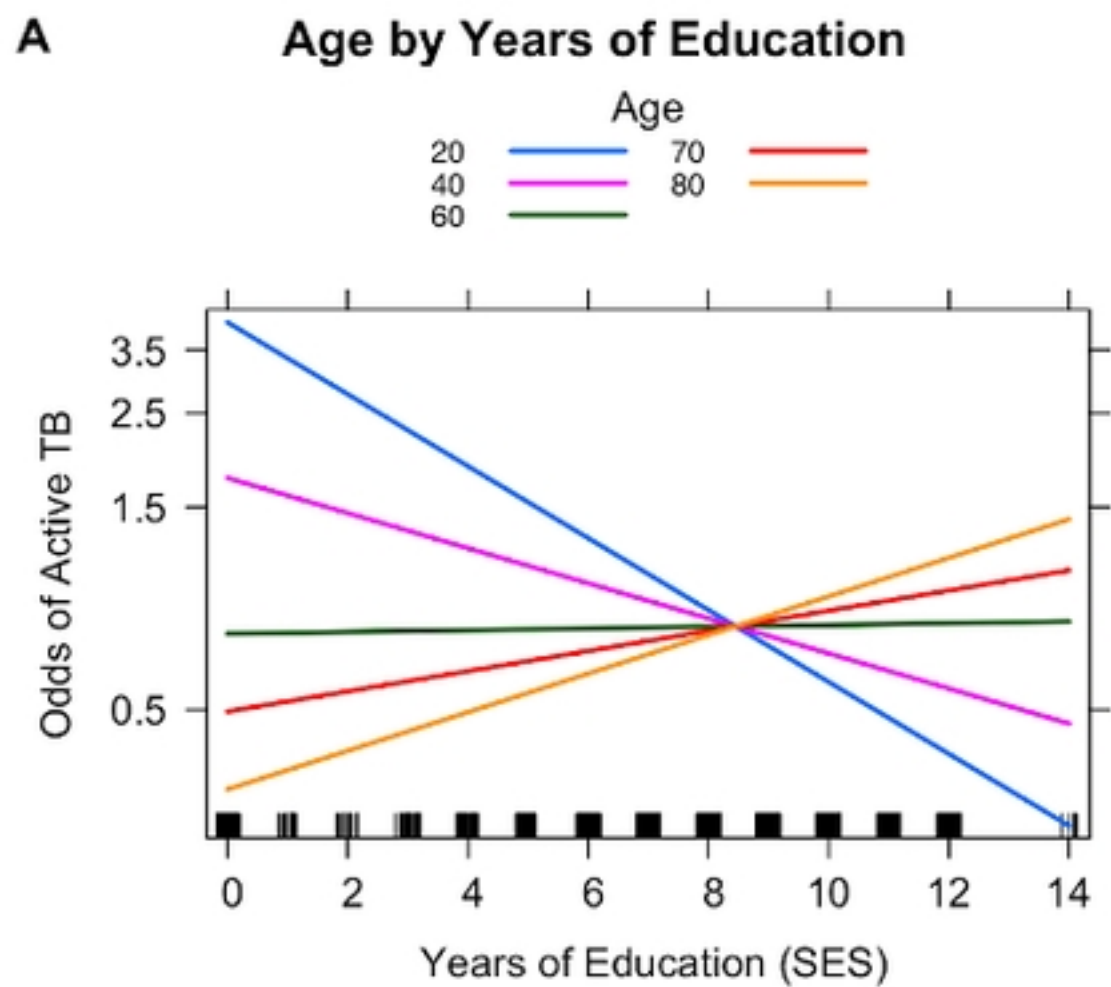


Figure 5

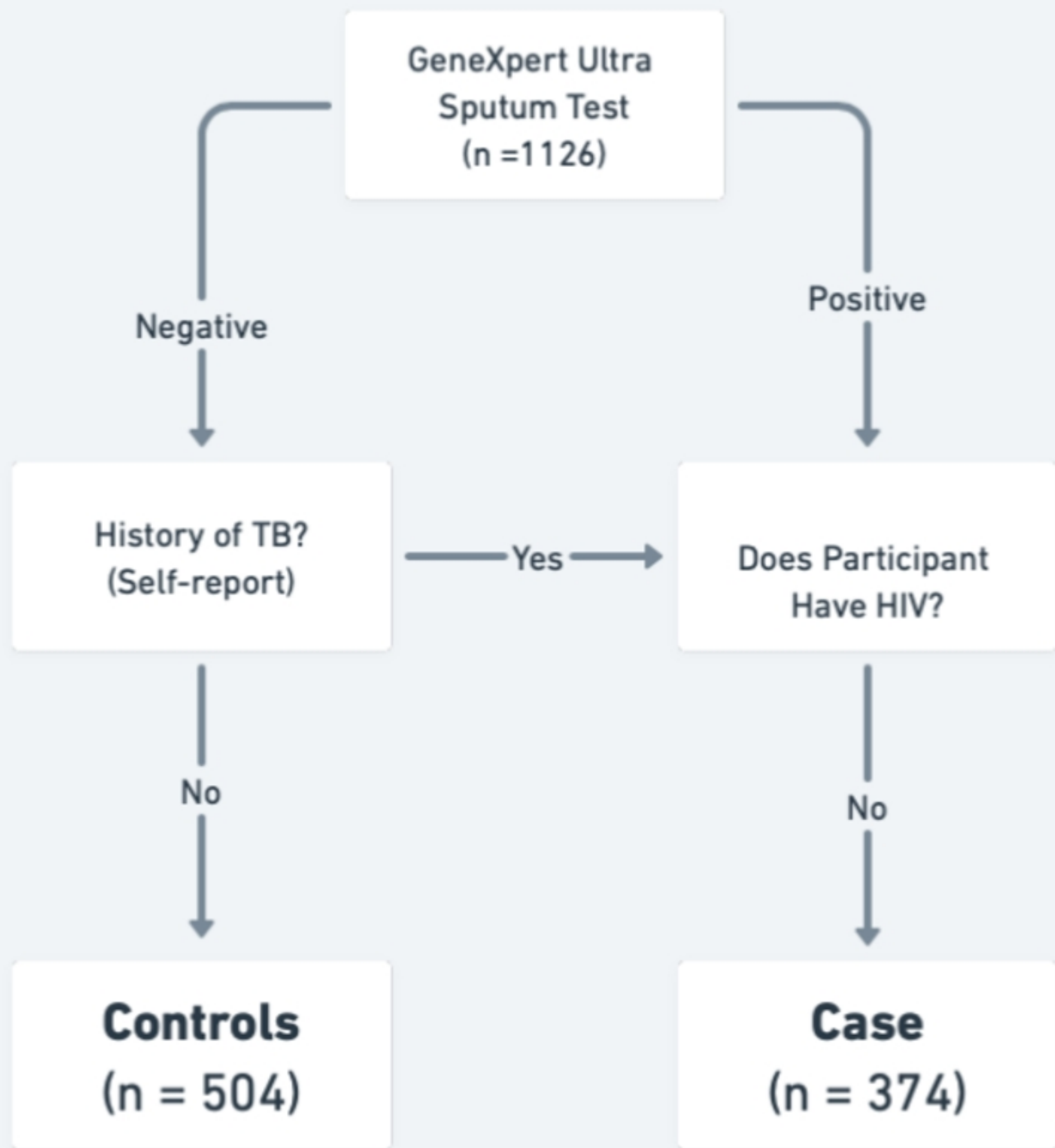


Figure 1