Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ 04/22/2019

Kellan Burrell Date

A county-level, ecologic analysis of correlations between recent tuberculosis transmission and socioeconomic indicators of poverty.

By

Kellan Burrell
Master of Public Health

Epidemiology

_____

Benjamin Silk
Committee Chair

_____

Yan Sun
Committee Chair

A county-level, ecologic analysis of correlations between recent tuberculosis transmission and socioeconomic indicators of poverty.

By

Kellan Burrell

B.S., Young Harris College, 2015

Thesis Committee Chair: Benjamin Silk, PhD; Yan Sun, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2019

Abstract

A county-level, ecologic analysis of correlations between recent tuberculosis transmission and socioeconomic indicators of poverty.

By Kellan Burrell

We sought to determine if recent transmission (RT) of tuberculosis (TB) was associated with socioeconomic or other ecologic factors. Identification of area-level factors associated with RT of TB disease could help better predict geographic areas at risk for TB outbreaks and allow investigators to better understand TB transmission. Ecologic analyses involving TB transmission have rarely been published in the past due to the difficulty of estimating RT. We used the plausible source-case method developed by France et al which integrates genetic, geographic, and epidemiologic data to determine if a case has a plausible source case. Data was pulled from the American Community Survey (ACS) tables as well as from the National Tuberculosis Surveillance System (NTSS) and logistic modeling was used to evaluate associations. Using this measure of RT, we found associations between poverty, black race, Hispanic ethnicity, and crowding. This study is the first study, to date, to use the plausible source-case method as a measure of RT to evaluate county-level factors and their association with estimated RT. Pediatric cases were underrepresented in this study as a consequence of the criteria within the plausible source case method and further research is needed to evaluate if inclusion of pediatric cases would change these associations. Further study using multilevel models that integrate area-level and patient-level characteristics would allow for even greater understanding of the associations between SES, demographics, and risk factors for RT of TB. Data sources for this study are regularly produced/updated and readily available for research which leads to this study being easy to repeat or modify in order to continual evaluate the associations between SES and RT at the county level.

A county-level, ecologic analysis of correlations between recent tuberculosis transmission and socioeconomic indicators of poverty.

By

Kellan Burrell

B.S., Young Harris College, 2015

Thesis Committee Chair: Benjamin Silk, PhD; Yan Sun, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2019

Table of Contents

Chapter 1: Background

Tuberculosis (TB) History, Evolution, and Biology

Tuberculosis disease is caused by members of the *Mycobacterium tuberculosis* complex (MTBC), which includes *M. tuberculosis, M. bovis, M. caprae, M. pinnepedii, and M. africanum*(1). *M. tuberculosis* is the most common form of TB in humans and most of the other members of the MTBC cause tuberculosis in both wild and domesticated animals (1). The exception being *M. africanum* which causes TB infection in humans in some regions of Africa. Infection with *M. tuberculosis (Mtb)* is still the primary source of TB cases in those regions with *M. africanum* only accounting for 38% of smear positive cases (2).

TB disease originated in Africa nearly 70,000 years ago and has spread from Africa following human migration patterns (3). Evolutionary pressures caused the disease to develop to survive in low density populations through long latent periods (4). The introduction of agriculture and domestication of agricultural animals led to increased population density and selection of strains that were more virulent and transmissible (3).

The bacteria itself is slow growing and possesses a strong cell membrane structure that is largely impermeable to drugs or extracellular compounds (1). The cellular membrane structure of *Mtb* is like other gram-negative bacteria with an asymmetric lipid bilayer with an outer layer composed of glycolipids and other waxy components and an inner layer of long fatty acids (1). Drugs that are effective in treatment of TB often target the synthesis of these cellular membrane components (1).

The characteristic of *Mtb* that has allowed the bacteria to persist for so long despite control efforts is the ability of the bacteria to go dormant. When faced with an infection the body mounts an immune response utilizing both granulomas, a type of cellular quarantine, and macrophages to

eliminate infection by taking up bacterium and digesting them. TB bacteria can sense the environmental conditions both within granulomas and macrophages and trigger a response that causes the bacteria to go dormant (1, 5). This dormant state stops replication, slows cellular metabolism, and activates an anaerobic metabolic pathway in order to survive for a long period of time (1, 5). The drastic change in the specific cellular mechanisms between active TB and dormant TB (also known as latent TB) make it difficult to control the spread of the disease. An individual that has never shown symptoms of TB disease can have latent infection and live with the pathogen for many years before some event triggers activation of the latent infection.

Active versus Latent TB Infection

The distinction between active and latent TB infection is critical to public health practice (6). Latent Tuberculosis Infection (LTBI) is often monitored in certain populations that are at a high risk for activation of the infection. These groups can be HIV infected individuals, individuals from countries with high endemic TB prevalence, or healthcare workers just to name a few (7). Identifying individuals with LTBI allows clinicians to administer treatment before the patient presents symptoms of the disease and before they are infectious (8). This control method prevents active transmission of disease and outbreaks.

Once an individual has progressed from LTBI to active disease the strategy for prevention becomes much more complicated and costly (9). It is estimated that ~10% of LTBI infections progress to active TB (10). However, control efforts must assume that every person in contact with a patient has contracted the disease (10). This means that a network of contacts is formed around each reported case of active TB and a contact investigation is carried out (12). These thorough and costly (13) investigations seek to identify any individual that had close contact with a TB patient and test them for the disease. A positive TB test for a contact adds them to the

transmission network and they are referred for treatment based on if they have either LTBI or active infection. The overall cost to control TB is lower if preventative screenings can identify LTBI before it progresses into an active and infectious form of disease (9).

TB Control Strategy

This unique transmission dynamic leads to a multi-faceted control effort by public health professionals. The most pressing and direct efforts are to identify all active TB cases and administer treatment to the sick individuals (14). After identification and treatment of infected individuals the next priority is to conduct contact investigations to identify any exposed individuals (14). After control of active cases and identification of contacts, the final strategy is to conduct targeted testing/screening to identify LTBI in at-risk groups and then refer those LTBI positive individuals for treatment (14). As national surveillance systems for TB grow and capture more cases of TB, it becomes difficult to identify which clusters of disease may be related to recent transmission of TB disease and which cases may be activation of LTBI, TB acquired elsewhere, or reactivation of previous disease (15). This distinction between active or previous TB is important because efforts need to be focused at controlling active TB to prevent outbreaks (15).

Tuberculosis in the US and Abroad

Tuberculosis incidence in the United States of America (US) has been declining for decades (16). The largest proportion, >80%, of TB burden in the US is TB acquired elsewhere or reactivation of latent TB (17). This is a difficult form of TB to control and a large portion of cases from reactivation or TB acquired elsewhere are individuals from high-risk groups (17). The most effective method of control for TB acquired elsewhere and LTBI is targeted testing of the groups with highest incidence of TB from these sources. However, there are ethical concerns

with targeted testing that inevitably targets immigrant populations and individuals with other chronic comorbidities (18). Targeted testing of some groups can be costly, logistically difficult, and perceived as discrimination (18). It may be desirable to attempt to refine TB control in the ~20% of cases that come from recent transmission of active TB in scenarios where targeted testing to prevent reactivation is difficult. Previously it was difficult to monitor or control recent transmission of TB because of long latent periods (15). However, recent advances in TB genotyping have made it easier to identify transmission chains and the probability of TB cases being related to one another (15).

TB Genotyping: MIRU-VNTR, Spoligotyping, TB GIMS

The Centers for Disease Control and Prevention (CDC) established the National TB Genotyping System (NTGS) in 2004 to provide genotyping services to state and local programs (18). *Mtb* Genotyping uses a hybridization assay called spacer oligonucleotide typing (spoligotyping) to detect variability in direct repeat region of the *M. tuberculosis* DNA (19). The region contains several copies of a 36-base-pair sequence separated by unique spacer sequences (19). These sequences vary between TB strains and can be used to identify the different strains. Variable number of tandem repeat (VNTR) typing is used in conjunction with spoligotyping to further refine genetic relationships between *Mtb* strains. VNTR typing focuses on candidate segments of the DNA segment called Mycobacterial Interspersed Repetitive Units (MIRUs) that contain tandem repeated sequences (19). The *Mtb* genome has a total of 41 MIRU loci but only 12 are selected for MIRU-VNTR typing (19). A TB case is assigned GENType designation based on the results of these two analyses (18). In 2010 the TB Genotyping Information Management System (TB GIMS) was launched as a secure, semi-automated, web-based system to track and monitor TB cases based on results of genotyping. This system, coupled with geospatial analysis

and algorithms built into TB GIMS, allows for outbreak detection based on patterns of TB cases with matching GENTypes (18). Currently the system identifies higher than expected concentrations of a TB GENType in a specific county compared to the national distribution of that genotype (18). A log-likelihood ratio (LLR) is calculated and the higher the ratio then the greater the possibility of a recent transmission event (18).

These methods of genotyping only consider a small percentage of the TB genome. It is possible, especially in areas with endemic strains of TB, for cases to have the same GENType but not be related through any transmission links. Whole genome sequencing (WGS) covers most of the TB genome and can be more readily used to identify recent transmission via genetic similarity (20). Simply put, if two cases in the same geographic area have TB with only a small number of single nucleotide polymorphisms (SNPs) difference in the genetic material of the two cultures then it is highly unlikely that the two cases are unrelated. Automated methods of integrating this WGS analysis are still underway but the technology is currently used to make prioritization decision in outbreak response at CDC (20). Public health professionals are trying to refine meaningful SNP cutoffs for identifying RT and there are new technologies in development to automate some of the analysis of WGS data (20). New technology aims to refine the GENType procedure and create automatically generated ‘strains’ that are less likely to be unrelated than the GENType designations currently are. This new technology is whole-genome multi-locus sequence typing (wgMLST) and will be used in tandem with whole-genome SNP sequencing (wgSNP) (20). Either of these technologies should be used in conjunction with traditional epidemiologic data to identify clusters of closely related cases (20).

TB Genotyping technology and Recent Transmission

Recent transmission is at the center of most TB control in the US. The importance of identifying recent transmission and the efforts taken to develop technology that aids in the identification of cases that may be related through transmission chains. The culmination of the efforts to automate identification of recent transmission was the development of the plausible source case method also known as the recent transmission (RT) algorithm (21). RT as defined by the plausible source case method will be considered separately from recent transmission as a concept for the remainder of this paper. Until development of the algorithm by France et al in 2015 there was a disconnect between genotype-based estimates of RT and field-based epidemiologic estimates of RT. A method integrating epidemiologic data and genotype-based approaches was developed. The approach used a set of criteria to determine whether a plausible-source case could be identified for each case in the study period. A plausible-source case must involve a respiratory form of TB in a patient over 4 years of age, be diagnosed within two years of the case under evaluation, resided in the same geographical area, and have the same RFLP pattern as the case under evaluation (21). In the event a plausible-source case was identified, each case was given the designation of attributable to RT (21). There were exclusion criteria placed on recent arrivals to the US even if a plausible-source case was identified (21). This method was tested against both field-validated methods as a gold standard and the current genotype-based methods for identifying RT.

The RT algorithm proved to perform as well, if not slightly better, than either field-based methods or genotype-based methods alone and overcame limitations of previous RT estimation approaches (21). Most notably the RT algorithm doesn’t assume that the first case in a cluster is the source case for the rest of a cluster, which is a weakness of the n-1 methods used prior to

development of the algorithm that used RFLP patterns to draw transmission chains from the first case of a cluster by the reported date as a source case for all others (21).

Applications of the RT algorithm

One aspect of the utility of the RT algorithm is the ability to modify the algorithm based on study needs or the environment in which the algorithm is being applied (21). Since first being developed the RT algorithm has been refined and used to describe RT trends based on individual risk factors and demographics of TB cases within the US (22). The goal of understanding these factors is to identify populations at higher risk for RT and guide control efforts (22). These types of studies have been done in the past but only using genotypic clustering which is not always indicative of transmission links (23, 24). Yuen et al. used the plausible source case method to analyze data collected from US National Tuberculosis Surveillance System (NTSS) and NTGS (22). Data was pulled for all cases reported between January 2011 and December 2014 and the RT algorithm was applied (22). Social network analysis was used to group transmission clusters (22). Findings from the Yuen et al. study showed that RT does not follow incidence trends. Only 9% of the variance in RT was explained by variation in incidence and five of the eight lowest incidence states had counties with more than 20% of cases being attributed to RT (22). These results illustrated the difference between high incidence areas with endemic strains of TB that present as isolated reactivation cases and areas where TB outbreaks occur in an otherwise low incidence environment. Individual-level characteristics most strongly associated with RT were less than or equal to 4 years of age (Prevalence Ratio (PR) = 5.1, 95% Confidence Interval (CI) 4.4-6.0), American Indian/Alaskan Native race (PR = 6.4, 95% CI 5.1-8.0), and homelessness (PR = 5.7, 95% CI 5.1-6.3) (22). A surprising result was that non-US born individuals had a strong negative association with recent transmission (PR = 0.3, 95% CI 0.3-0.4) (22). Non-US

born individuals are a group that have shown a much higher than average incidence of TB disease. A protective effect of being non-US born on RT is further evidence of RT trends not following established trends in TB incidence. Overall 14% of genotyped cases in the US between 2012-2014 were attributable to RT (22). This is lower than the 22% of cases that are clustered genotypically which supports claims that genotyped-based clustering alone is insufficient for accurately identifying RT (22). The Yuen et al study makes a compelling case for recent transmission being concentrated in US born populations and most TB in non-US born populations being associated with TB incidence in their country of origin (22).

Ecologic Studies before development of the RT algorithm

Genotyped-based clustering alone has been used in the past as a proxy measure for RT despite the inaccuracies of genotype-based clustering as a RT measure. Despite the poor performance of genotype-based methods alone, a study by Oren et al found associations with county-level socioeconomic status (SES) measures (23). The study used data from 2004-2008 in King County, Washington to draw conclusions about the relationship between tuberculosis transmission and socioeconomic status at the block group level (23). The study used socioeconomic positioning (SEP) scores as their block group level measure of SES and found that when considered in a multilevel model with individual characteristics, the effect of SEP at the block group level attenuated individual risk factors, making them have a smaller association with risk of genotypic clustering (23). However, this study used a method of genotypic clustering like the GENTyping procedure discussed earlier. This method of genotypic clustering is prone to overestimation of recent transmission and attributing cases to clusters when they may be genetically diverse using more thorough methods. This is especially true when considering small geographic areas like a single county where there may be endemic GENTypes that would, on the

surface, seem to cluster but are activation of latent disease and not recent transmission. Despite these limitations the study found that block group levels with lower SES exhibited greater odds of genotypic clustering (23).

A study by Myers et al also evaluated the association of county level SES factors on recent transmission of TB (24). The study used pediatric cases as a proxy for recent transmission because nearly all clusters linked to a case of pediatric TB involve recent transmission. The long latent period for reactivation cases means that it is unlikely for a case of active TB in a pediatric patient to not be recent transmission from another active TB case in an adult. This method of categorizing cases as recent transmission results in a small number of false positives but misses many clusters of transmission that don't include a pediatric case. TB case data for 10 years between January 1, 1993 to December 31, 2002 were collected and geocoded for the analysis (24). There were 3208 cases of TB in the pediatric population during the years included in the study resulting in a crude incidence of 4.1 cases per 100,000 person-years. Incidence rates varied between census tracts from 0 to 230 per 100,000 person-years. The study found that census tracts with lower median incomes, more Black individuals, and census tracts with more non-US born individuals have more new tuberculosis transmission (24). We would expect similar results in a nationwide study using a more accurate measure of recent transmission if pediatric TB cases are a good proxy for recent transmission. However, there may be important characteristics of TB transmission in clusters without pediatric TB cases that aren't captured by a study focusing on that subset of the population.

The link between socioeconomic status (SES) and TB

Socioeconomic status characteristics can be risk factors for many diseases, TB is no exception with transmission being linked to lower income groups (25). Olson et al in 2012 found

significant associations between education, crowding, income, and unemployment among US-born individuals in a dataset comprised of all reported TB cases between 1996 and 2005 (26). The study also found that the associations with SES and TB rates are lower and only crowding and income are significantly associated (26). This difference may be due to the overall higher rates of TB in non-US born populations which results in observing a smaller association with SES and TB rates overall. The study concluded that TB rates in non-US born populations are influence more by the experiences individuals had in their country of origin than experiences in the US (26). This finding aligns with findings in previous studies that show higher rates of RT in US-born individuals than non-US born individuals (22, 27).

Summary of TB recent transmission research to date

There have been several studies that evaluated SES factors related to TB rates and transmission (23, 24, 25, 27) and few studies that have evaluated recent transmission of tuberculosis based either on SES factors or individual risk factors (22, 23, 24). Despite studies finding significant associations between several individual risk factors, SES factors, and RT/clustering of TB, there have been no studies to date that use the CDC RT algorithm to evaluate associations between area-based SES factors and transmission of TB. The RT algorithm has potential to provide more accurate estimates of the significance and magnitude of associations between TB transmission and area-based SES factors. This understanding can allow public health professionals to have better situational awareness of TB in the US as well as better describe outbreaks that happen outside of an area that has higher TB incidence.

Chapter 2: Introduction

Tuberculosis (TB) prevention and control is centered around outbreak investigation and identification of ongoing recent transmission (RT) of TB. TB exists in three main forms in the US: active disease, latent disease, and TB acquired elsewhere. TB control often lumps latent TB and TB acquired elsewhere into one category of disease that is considered disease from reactivation. Most of the TB in the US is reactivation or TB acquired elsewhere and discerning cases of active TB from transmission that occurred years or even decades ago is critical for directing resources towards stopping active transmission (22). Therefore, the concept of identifying RT is important to control efforts. TB acquired elsewhere is difficult to prevent and control with targeted testing of high-risk groups being one of the only effective interventions for prevention. Active TB, on the other hand, can be controlled and managed via contact investigations, active case finding, and cluster investigation. These control methods, however, are costly and conducting contact investigations around cases that appear to be clustered but are in fact TB acquired elsewhere wastes precious resources. Understanding the factors both at the county and individual level that may increase risk of RT helps focus these resources.

Trends in TB incidence are well-documented by publications like the annual reported tuberculosis in the United States document. However, findings from previous studies evaluating RT of TB in the United States (US) shows geographical heterogeneity and poor prediction by TB incidence (22). Heavy focus is placed on previously known risk factors for TB disease when conducting outbreak investigation and casual risk assessment for TB incidence, but these factors are not always correlated with the outbreaks in TB disease that we see occurring throughout the country. We seek to determine if RT is associated with socioeconomic or other ecologic factors.

Identification of factors associated with RT of TB disease could help better predict geographic areas at risk for TB outbreaks and allow investigators to better understand TB transmission.

Ecologic analyses involving TB transmission have rarely been published due to the difficulty of estimating RT. Past studies have used either clustering of genotyped isolates (23) or pediatric cases (24) as indicators of RT. After these earlier studies, a more accurate and systematic method of attributing cases to RT nationwide have been developed (21). This new method accounts for some of the inaccuracies in genotype clustering as a method of identifying RT. Using pediatric cases as an indicator of RT avoids false negatives as it is well documented that pediatric TB is almost always due to direct transmission. However, transmission chains do not always contain a pediatric case and as such this method may miss a large proportion of RT clusters and under or over estimate the effect of ecologic factors. Genotyped-based clustering alone has been used in the past as a proxy measure for RT despite the inaccuracies of genotype-based clustering as a RT measure. Using strictly genotyping poses an entirely different set of limitations as we know that these methods alone do not correlate with RT (21) as accuracy of this method is largely dependent on the geographic unit and time windows over which a study is conducted. The approach to RT identification developed by the Division of TB Elimination (DTBE) at the Centers for Disease Control and Prevention (CDC) better accounts for some of these inconsistencies and allows better understanding of RT clusters and as such will be used as the method by which we define RT in this study.

The plausible source case method used in this study to estimate RT involves comparison of reported TB cases within the US. The method compares each reported case of TB with all other cases that involved a respiratory form of TB diagnosed 2 years before or 3 months after the given case in an individual greater than 4 years of age. The plausible source case must have also

resided within a set geographic distance of the given case. Genetic data originally was integrated into the algorithm via comparison of RFLP patterns but since the development of the algorithm genetic data is integrated via inclusion of more advanced genotyping data. The refining of genotype-based methods for identification of RT clusters is critical for improving control efforts and WGS is a critical component that avoids some of the false positives associated with other methods. Automated integration of WGS data is still being developed but we do currently have more accurate methods for integrating genetic data than was available when some other studies have been conducted. The RT algorithm in the current state has been shown to be more effective in estimating RT than other methods in the past (21, 22).

We hypothesize that there will be an association between RT as defined by the plausible source case method developed by France et al. and county-level factors such as income, employment, educational attainment, health insurance, crowded housing, and population density. These factors have been shown in previous literature (22-27) to be linked to TB transmission dynamics.

Methods

Data Sources and Data Cleaning

This study utilized data from the National Tuberculosis Surveillance System (NTSS) and the Annual Communities Survey (ACS) produced each year by the Census Bureau. Data from NTSS included patient level data on county, state, county FIPS codes, count date, variables on RT status (generated using an updated version of the original RT algorithm [Figure 1]), status variable for if the case has been genotyped, zip code, and GENType. After assigning plausible source cases across TB GENType clusters, the resulting case pairs are grouped into transmission clusters using PROC OPNET in SAS 9.4. This PROC is a social network analysis procedure that

links cases into clusters. This procedure is carried out regularly by statisticians and epidemiologists in DTBE as a part of the procedure for generating the genotyping section of the division annual incidence report. As such, the variable identifying a case as attributable to recent transmission is readily available in NTSS data available for use by guest researchers following proposal to the analytic steering committee for approval. Data on age, alcohol use, HIV status, homelessness, injection and non-injection drug use, race/ethnicity, sex, and country of origin were also obtained for secondary analyses but were deemed unnecessary and excluded from the final analysis dataset. ACS 1-year estimates were originally used as census documentation suggests these estimates over 5-year estimates when using economic variables that may change significantly year to year. However, 5-year estimates were also pulled and difference in variance between the 1 and 5-year estimates was negligible. The 1-year ACS estimates included data from only 822 counties for most variables and even though the 822 counties aligned with the counties in the final analysis dataset (only ~2% of counties missing for any single variable in the final analysis) the decision was made to use the more complete 5-year estimates which contained data for over 3,000 counties. The 1-year ACS estimates contain fewer counties as there are response level requirements for the census to publish data for a county and those response requirements are not met for the 1-year estimates resulting in the exclusion of many small counties. This can be observed by the average county population estimates for the entire US being 90,000 less using the 5-year estimates than when using the 1-year estimates (Tables 1&2).

The ACS data used in the analysis was pulled from 5 different subject tables. The tables and variables used are included in Table 3 together with the resulting variable names in the analysis dataset and descriptions of each data element.

Data was imported into SAS 9.4 for cleaning and analysis. SAS statements were used to pull two-digit state Federal Information Processing Standard (FIPS) codes to be concatenated with 3-digit county FIPS codes from the NTSS dataset. FIPS codes were used as a standard way to identify county and state between NTSS and ACS data and to account for any possible inconsistencies in county/state coding between the two datasets. NTSS data from 2015 – 2017 was aggregated to the county level using PROC FREQ to generate counts of TB cases per 5-digit FIPS code area stratified by RT status. The resulting dataset were then merged, and a proportion of RT was calculated from counts of genotyped cases eligible to be assessed for RT (denominators) and counts of cases attributed to RT (numerators) in each county. Any county with less than 10 total genotyped cases eligible to be assessed for RT was excluded from the analysis. This cutoff was to prevent counties with low case counts from skewing the distribution of RT proportions. PROC UNIVARIATE was used to find the overall median of the resulting proportions and that median was used as a cutoff for establishing a county as above or below median RT. The decision to dichotomize RT was made in order to establish a method to designate counties as having a higher proportion of RT without making the decision arbitrary. In previous publications, specifically the DTBE annual report, the percentage RT used for a cutoff was 14%. This percentage, however, was an arbitrary cutoff and we desired a method for categorization that could be repeated and yield a similar dichotomization of data. An arbitrary cutoff would not be affected by overall changes in incidence and transmission which would prevent the cutoff from capturing the yearly changes in TB transmission dynamics within the US.

ACS data from 2016 was pulled from the American Fact Finder advanced search tool. The data was imported to SAS 9.4, variables renamed, and then merged together before being merged

with the aggregated NTSS data. Data missingness was limited due to the standardization and completeness of both ACS and NTSS datasets.

Modeling Strategy and Model Selection

Logistic regression was used for all models in the analysis. The outcome event was coded as a county being above the median proportion RT and each predictor was first considered alone in a univariate model to get univariate associations (Table 4).

The multivariate model selection process started with a fully parameterized model containing 44 predictors which included every predictor listed in table 3 as well as interaction terms to account for effect modification based on combinations race/ethnicity, employment status, country of origin, and income. Many of the parameters could realistically be directly related so care was taken to do thorough collinearity diagnostics. Condition indices (CNI's) and variance decomposition proportions (VDP's) were generated using a collinearity diagnostics SAS macro. CNI's greater than 30 with 2 or more VDP's greater than 0.5 indicated collinearity issues. The fully parameterized model was assessed and after each generation of CNI's and VDP's the variable with the highest VDP was dropped (care was taken to retain a hierarchically well formulated model) and diagnostics were run again. This process was conducted iteratively until there was no longer a collinearity issue. The resulting model contained parameters for proportion Hispanic ethnicity, proportion Black race, proportion reporting Asian race, proportion non-US born, more than 1 occupant per room, proportion unemployed, proportion of individuals without health insurance, proportion of families below poverty level, and proportion of individuals with a bachelor's degree or lower. Interaction assessment was conducted after collinearity assessment and no interaction terms were significant and as such were dropped from the model.

After reviewing results from the full multivariate model, we decided to construct a parsimonious model containing only the significant and borderline significant parameters. This model contained proportion of families below the poverty level, proportion reporting black race, proportion reporting Hispanic ethnicity, and the crowding variable. We used Hosmer and Lemeshow goodness of fit tests to evaluate fit there was no evidence for lack of fit in either the full or the reduced parsimonious model. We also looked at the ROC curve for each model and found that both models had c statistics over 0.7 indicating good models. Finally, to test if the dropped terms were significant to the predictive value of the model, we ran a likelihood ratio test and found that, with a large p-value of 0.3, the dropped terms were not significant to the overall model. We have reported both the full multivariate and the reduced parsimonious model, but final conclusions are made using the parsimonious model with 4 predictors.

Each parameter in the model is a proportion or percentage of the population in a county reporting the parameter. Logistic regression produces parameter estimates that, when exponentiated, represent the change in odds based on a 1-unit increase in the predictor. This result is not intuitive to interpret and as such a unit change for each parameter was derived to better portray the results. The range of each parameter was split into quintiles. The quintile bin widths for parameters ranged from 8% to 20% except for proportion of households with more than 1 occupant per room and the proportion unemployed which were 3% and 6.8% respectively. We decided that 10% was an acceptable unit change for all parameters in the 8 to 20% range and would be less confusing to present than a specific unit change for each parameter. This unit was too large for crowding and unemployment parameters so 5% was selected as a unit of change for those parameters. A unit's statement was added to the logistic regression procedure in SAS

which automatically reported odds ratios and confidence intervals in the specified unit for each parameter.

We attempted to stratify data by some measure of population density as there were concerns that the distributions of certain factors may differ significantly between urban and rural areas. However, none of the delineation measures we attempted to integrate were sufficient in splitting suburban areas from rural areas. This will be further addressed in the strengths/limitations section.

Results

Descriptive Statistics

Table 1 and 2 contain simple descriptive statistics of included counties, counties above median RT, and counties below median RT compared to overall county-level factors are of interest in this study. The median proportion RT that resulted from our categorization was 9.1%. The resulting dataset contained 280 counties with more than 10 genotyped cases eligible to be assessed for RT with 140 being categorized as above median RT and 140 categorized as less than median RT. We see that counties included in the study (counties with more than 10 genotyped cases eligible for assessment for RT) are larger, on average, than counties overall in the US. Furthermore, we see that counties with more RT are slightly more populated than those below median RT. We see a very different distribution of race/ethnicity in included counties than we see nationally and distribution of place of birth in included counties is different from the national distribution. We see almost 3 times as many non-US born individuals in included counties as we see nationwide. It should be noted, however, that the percentages represent the average proportion of individuals in each county reporting a certain parameter, not the distribution of TB

cases in the US which would show a majority non-US born instead of the ~14% in table 1 and 2. Unemployment is higher in included counties and higher among counties with more TB than in counties with below median RT. Income is higher in included counties compared to nationally although poverty is similar to national levels.

Statistical testing for difference in means was not applied to data in table 1 and 2 as the estimates therein come from census survey data which has inherent variance and has already been subjected to statistical transformation. Applying statistical tests to this type of data was deemed inappropriate.

Univariate Analyses

Univariate analyses found significant associations between counties being above the median proportion RT and black race, crowding, unemployment, health insurance, education, and poverty. Poverty was the association with the greatest magnitude of effect in univariate analyses with an unadjusted odds ratio (OR = 5.0, 95% CI 2.79 – 8.98). Hispanic ethnicity, country of origin, and Asian race were not significantly associated with a county being above the median RT in univariate analyses.

Multivariate Analyses

Adjusted ORs (aORs) were attenuated when controlling for all other factors in the full multivariate analysis with poverty being the only statistically significant association (Table 4). It is worth noting that Hispanic ethnicity (aOR = 0.7, 95% CI 0.46 – 1.01) and black race (aOR = 1.4, 95% CI 0.99 – 1.88) were marginally significant (p-value < 0.1) in the full multivariate model. Crowding, to a lesser extent, was also an interesting finding to the researchers in that the

aOR was 3.1 and the CI spans above 17 on the far-right tail even though the left tail is at 0.54. This result is not statistically significant but suggests crowding can possibly have a large effect on the odds of elevated RT.

However, when we apply the parsimonious model to these data, we find that the exclusion of Asian race, proportion non-US born, proportion unemployed, and educational attainment results in significant associations with the remaining variables. Hispanic ethnicity (aOR = 0.7, 95% CI 0.52 – 0.91) exhibits a protective effect on a county having more than the median proportion RT and counties with a higher proportion reporting Black race (aOR = 1.3, 95% CI 1.01 – 1.74) are 30% more likely to have more than the median proportion RT. Crowding (aOR = 6.2, 95% CI 1.65 – 23.16) has the largest effect and the widest confidence interval which suggests a large amount of variability which could represent some counties having more crowding but less RT than would be expected. Poverty (aOR = 4.4, 95% CI 1.97 – 10.01) also has a large effect and this result aligns with a priori assumptions about RT of TB.

Discussion

We conducted one of the first studies, to date, evaluating associations between TB recent transmission and county-level SES factors. We found that, when controlling for race/ethnicity, and crowding counties with above median RT have 4.4 times the odds (95% CI1.96 – 10.01) of 10% more families below the poverty level. We also found a protective effect of counties with more Hispanic individuals (aOR = 0.7, 95% CI 0.52 – 0.91). Black race had a significant effect (aOR = 1.3, 95% CI 1.01 – 1.74) and crowding had a large effect (aOR = 6.2, 95% CI 1.65 – 23.16) and a CI that extended above 23.16 on the far-right tail. We saw, overall, attenuation of all factors when going from univariate models to the full multivariate model. This is consistent

with results in previous studies (23-27) that found factors together have a smaller effect than alone. Median income was dropped from the original model during collinearity assessment in favor of retaining poverty as a parameter. Poverty is coincident with various other TB risk factors as areas with a larger proportion of families below the poverty level can often have more uninsured, more minority populations, less access to health care, crowded housing, substance abuse, and homelessness. This concentration of risk factors can drive transmission of TB and delay identification of active TB cases as people don't seek treatment until they have exposed many other individuals.

Country of origin was controlled for in the full model analysis analysis which would indicate that poverty is a significant factor in TB transmission regardless of if the county has a high proportion of non-US born individuals or not. Individual level studies using the RT algorithm found that RT is more common in US-Born populations (22). These two findings together suggest that counties with large populations of poor, US-Born individuals may be at a higher risk overall for outbreaks of TB clusters resulting from RT. Hispanic ethnicity in the multivariate model resulted in a protective effect (aOR = 0.7, 95% CI 0.52 – 0.91) where in the univariate model we saw a null effect (OR = 1.0 95% CI 0.89 – 1.14). This may be due to controlling for counties with a large immigrant population in the multivariate analysis. Cases of TB among Hispanic individuals are often from genotypes endemic to Mexico and are present in US-Mexico border states. What may be captured in the protective effect from the multivariate model is the clustering of Hispanic TB cases in counties with large Hispanic populations where there is very little TB transmission. Hispanic ethnicity spreads throughout the US so the variable alone wasn't enough to capture the effect of areas in border states where TB incidence is tied TB acquired in Mexico. Even among US-Born individuals in these border states TB incidence may not indicate

RT because of long term exposures to non-US born family members from Mexico. The DTBE annual report publishes maps showing the spread of RT throughout the US and we can see that there are only a few counties along the US-Mexico border that have high levels of RT which would support our hypothesis on the root cause of a protective effect of Hispanic ethnicity. However, this matter warrants further investigation to truly understand the factors involved in the protective effect. Black race is statistically significant at an alpha of 0.05 (aOR = 1.3, 95% CI 1.01 – 1.74). This result demonstrates that there is some effect of the proportion of residents reporting black as their race in a county on RT. In the US Black race often goes hand in hand with these poor communities described earlier where risk factors concentrate and access to medical care is either scarce or too expensive to take advantage of. The association with crowded housing was significant in univariate analysis but is not statistically significant when controlling for other factors. However, the association is significant in the final parsimonious model where we drop country of origin along with several other factors. This possibly represents an interesting dichotomy of crowding in the US between immigrant populations and US-Born populations. Immigrant populations often are in crowded housing situations immediately after arrival and there are communities across the country where refugees and other immigrants get placed or migrate to after entering the country. Despite crowded housing in those areas, RT is not common because of the dynamics of recent arrivals being involved in a recent transmission chain. This means that many counties may have a high proportion of households with more than one occupant per room but not have above median RT because the immigrant populations living in crowded housing haven’t contributed RT cases to the numerator for that proportion. On the other hand, crowded housing among US-Born individuals often coincides with individual level TB risk factors like poverty and homelessness. Crowded housing in US-Born communities is also likely

long-term crowding compared to crowded housing in immigration hubs where the individuals often move into less crowded living situations once they are better established in the country or leave the country altogether. These two extremes could be what is driving this large confidence interval for crowded housing when controlling for other parameters such as country of origin. This result, nevertheless, should not be ignored as we can see that, despite the wide confidence interval, there is some link with RT and crowded housing that is drawing the interval up above 17 and we see that in the final parsimonious model the association is significant. This study was effective in confirming the relationship between poverty and TB RT. We can clearly see the impact that county-level poverty has on RT and even a small 10% increase in county-level poverty can increase the odds of a county being above the median RT by over 4 times. The study is also readily repeatable which should enable future research to account for changes in the overall demographic or economic characteristics of the country and TB incidence in the country. Both data sources used for this study are regularly collected and available to researchers for analysis. This data availability is a strength that is not found in studies using data collected from clinical data or under special data use agreements that can't necessarily be repeated on-demand year after year.

There are several weaknesses to this study design, mostly with the nature of county-level ecologic data. The biggest weakness in this study is the inability to accurately account for the variation of both within county demographics and between county demographics. Counties within the US vary widely both geographically and demographically and this is difficult to control for given the data we have available. Urban counties have more variability in economic and demographic profiles than rural or suburban counties. One can see this in the through the elevated median income of included counties compared to national averages in table 2. Counties

in this study were mostly urban counties and a large portion of the dataset was comprised of counties surrounding large population centers. This type of environment means that even an outlier resistant measure of center like the median can be elevated as there is a large portion of very wealthy individuals in the population of urban counties while there are still an above average number of families living below the poverty level. This is often observed as an apparent lack of a middle class in urban areas where there is definite split between a large population of wealthy individuals and a large population of very poor individuals. This dynamic of economic status in urban counties is difficult to account. Beyond within county variation there is diversity across the nation between different counties. A county at the city center in Idaho, for example, is very different than a county at the city center in California. There are differences both in the economic profile and the geographic spread of counties across the country. Counties on the west coast and in the middle of the country tend to be larger and sprawl across large geographic areas resulting in extreme variability. Comparatively counties on the east coast are smaller and often more concentrated both in terms of population and demographic characteristics. The issue is really one of large diversity and variability in some areas of the country compared to limited diversity in other areas of the country. Ideally these factors could be controlled via matching or stratification, but the nature of the RT algorithm makes it difficult to apply at smaller geographic levels and there are limited ways to delineate county-level data into urban and suburban areas.

The design of the study itself is a limitation in that we are only account for county-level ecologic factors and are dichotomizing an outcome for the sake of applying logistic regression. Logistic regression is a powerful tool for analysis and provides easy to interpret results with minimal room for misinterpretation. However, there is some criticism that dichotomization of the outcome is arbitrary and could cause some associations to be missed depending on which

counties fall just on one side of the cut point or the other. The reason this limitation is not addressed, and another methodology used, is because the main goal of this study and the question of importance is really a question of extremes. Tuberculosis control in the US, specifically outbreak response and control of RT, is focused on managing the large ongoing outbreaks of recent transmission. The comparison we are concerned with is between counties with little to no RT but incident cases of TB and counties with a very large proportion of RT among their incident cases. Dichotomization of the outcome is enough to capture that comparison. The natural next step for this study is to integrate individual level data into the county-level analysis via hierarchical modeling. Considering individual-level and county-level factors together will provide the most complete picture of TB RT in the US.

Another future direction would be to attempt to account for the weakness of underrepresentation of pediatric cases in analyses using the RT algorithm. Accounting for this limitation may be as simple as adding any case under a certain age to the numerator of cases attributable to RT if they have not already been captured by field epidemiology linking cases or the algorithm. However, the issue of what age cut off to use and how adding these cases would affect the overall analysis needs to be evaluated.

Chapter 3: Summary

In this paper we investigated associations between county-level socioeconomic factors and recent transmission of TB as defined by a CDC-developed plausible source case method known as the RT algorithm. This algorithm integrates epidemiologic data, clinical data, and genetic data to identify plausible source cases for every reported case of TB over a specified time frame. The algorithm is run regularly across the entirety of NTSS data and helps public health professionals better understand how TB cases are related.

We used NTSS data and ACS 2016 5-year estimates of SES factors at the county-level to find associations between several SES factors and county-level RT above the median proportion RT in our study population. Poverty (aOR = 4.4, 95% CI 1.97 – 10.01), Hispanic ethnicity (aOR = 0.7, 95% CI 0.52 – 0.91), and Black race (aOR = 1.3, 95% CI 1.01 – 1.74) were significantly associated with RT above or below the median proportion RT.

Crowded housing was expected to be associated with RT of TB and was significant in univariate analysis, but the full multivariate model showed no statistically significant association with a p-value of 0.20. The final parsimonious multivariate model resulted in a significant association between crowding and median proportion RT (aOR = 6.2, 95% CI 1.65 – 23.16). Previous research has provided a possible explanation of the very wide confidence interval in the full multivariate analysis (95% CI 0.56 – 17.21) by describing an anomaly in crowded housing where immigrant populations often live in crowded housing but don’t contribute to elevated recent transmission.

Public Health Implications

TB control in the US is currently a multi-pronged approach. We described how TB control focuses on active TB transmission as well as identification and treatment of latent disease.

Identification of latent disease involves describing populations at a higher risk for latent infection and then targeting testing in those risk groups. However, we have seen that TB risk groups for LTBI do not align with groups at risk transmission of active, recently acquired disease. This study takes a step to further describe the populations or environments where recent transmission of active TB is taking place. This knowledge can help focus control efforts and strengthen cluster assessment and outbreak response.

We also described the cost of active versus latent TB and how contact investigations around active cases of TB account for a large portion of the budget and manpower of state and local TB partners. The heterogeneity of RT and independence of RT from incidence trends means that unexpected outbreaks can occur in low incidence states and counties. Funding for TB control programs is largely linked to incidence measures and as such may be insufficient to fund outbreak response in low incidence states and counties. This research helps public health professionals better understand how RT occurs independent of incidence and may lead to further studies that could better anticipate outbreaks in areas otherwise untouched by TB.

The percent decrease year to year in TB incidence in the US has started to plateau as we reach what may be an endemic level of TB incidence that can't easily be reduced. A large amount of TB transmission in the US is among immigrant populations that acquired TB elsewhere and became active cases after immigration. These cases have no readily identifiable plausible source case and other than screening all immigrants entering the country for LTBI it is difficult to completely prevent these cases. However, we can reduce the incidence of cases resulting from recent transmission of active TB by strengthening our TB control programs. This research provides understanding that is pivotal to informing local, state, and federal TB control professionals about the specific environments that increase odds of recent transmission.

Possible Future Directions

The next step for this project is to refine the model and explore other distributions and categorizations of RT as an outcome. Work will continue and new models and distributions will be integrated into the analysis. A large part of the continuing research on this topic will be the development of a multilevel/hierarchical model that integrates the patient level surveillance data with these county-level counts and proportions.

Pediatric cases will be better represented in future analyses as the final goal is to provide as complete a picture of TB recent transmission in the US as possible. There are numerous ways to include pediatric cases and different approaches from simply adding cases that meet a certain criterion all the way to modifying the plausible source case method need to be explored.

Future studies should also attempt to control for the diversity in counties and between counties. We described, at length, the issues with the current analysis and county-level diversity. This limitation may not be able to be addressed without accessing other datasets or conducting a more in-depth study that utilizes more robust datasets than the surveillance datasets we used in the analysis.

We believe there is potential for this research to be integrated into regular surveillance if the models are further developed and find even more informative results. All data used in these analyses are regularly collected and reported in a clean and standardized format. Minimal work would be required to regularly reproduce updated estimates of these area-level associations with RT and looking at these associations regularly may provide a better understanding of transmission dynamics and how RT changes over time.

Figure 1. Plausible Source-Case Method for Attributing Tuberculosis Cases to Recent Transmission (RT)

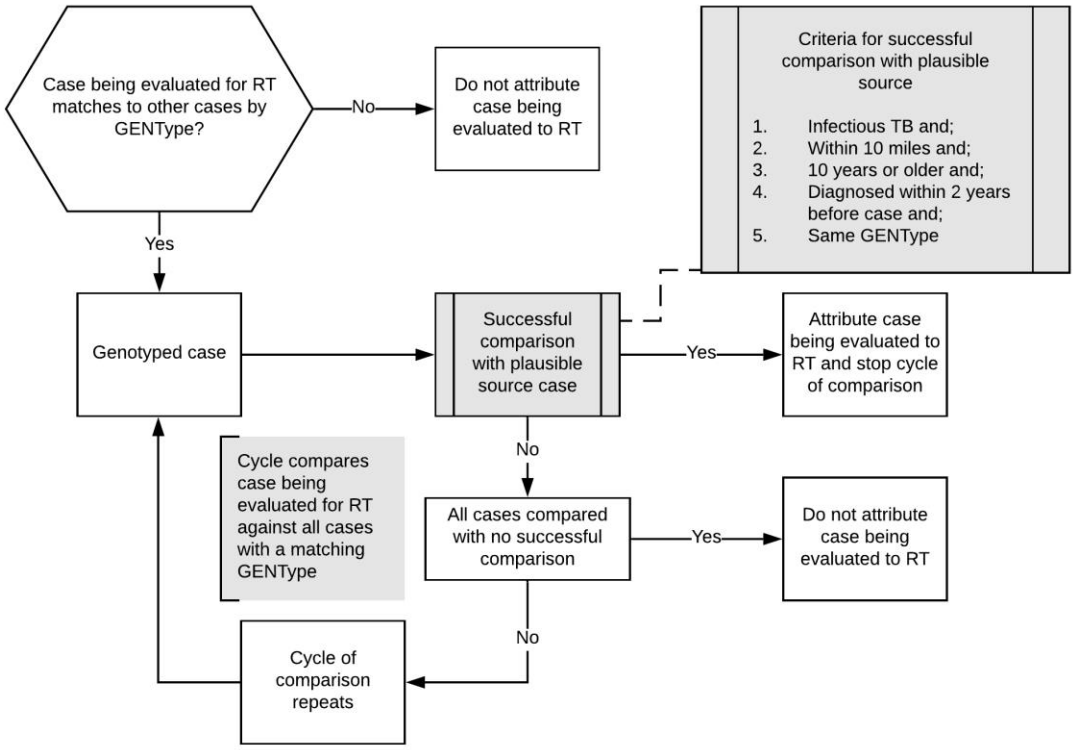

Table 1. County-Level Census Measures from 1-Year Estimates								
Characteristic	**Included Counties**		**Above Median RT**		**Below Median RT**		**Overall United States**	
Population	**Est[1]**	**STDEV[2]**	**Est**	**STDEV**	**Est**	**STDEV**	**Est**	**STDEV**
2016 ACS Estimate	711302	891526	805093	1156488	622102	517767	191756	460507
Race/Ethnicity[3]	**Mean%**	**STDEV**	**Mean%**	**STDEV**	**Mean%**	**STDEV**	**Mean%**	**STDEV**
White	68.3	15.4	64.1	17.0	72.4	12.4	78.2	15.1
Black	16.2	14.4	20.1	16.3	12.5	11.1	11.0	12.5
Asian	5.9	6.1	6.1	7.5	5.8	4.5	11.7	13.5
Hispanic	18.5	17.3	19.3	19.5	17.7	15.0	3.4	4.4
Place of Birth	**%**		**%**		**%**		**%**	
Non-US Born	14.3	9.1	14.4	10.5	14.2	7.5	8.2	7.3
Occupants Per Room	**%**		**%**		**%**		**%**	
More than 1	3.5	2.6	3.9	2.9	3.1	2.2	3.0	2.2
Employment	**%**		**%**		**%**		**%**	
Unemployed	3.7	1.0	3.8	1.0	3.6	0.9	3.5	1.1
Employed	60.1	5.7	59.1	5.8	61.1	5.5	58.4	6.5
Income	**USD ($)**		**USD ($)**		**USD ($)**		**USD ($)**	
Median Income	62153	17675	57108	15568	67014	18267	58111	15303
Health Insurance	**%**		**%**		**%**		**%**	
insurance	85.7	9.1	85.6	10.5	85.8	7.5	91.8	7.3
no insurance	8.7	4.3	9.7	4.7	7.8	3.7	8.0	4.0
Proportion of Families below poverty level	10.3	4.8	11.8	5.0	8.8	4.0	9.6	4.5
Educational Attainment	**%**		**%**		**%**		**%**	
Proportion Bachelors or Higher	33.8	10.9	31.6	10.6	35.9	10.9	29.3	10.5
Proportion High School Graduate	87.8	5.6	86.4	5.9	89.2	5.0	88.6	5.0

[1] Estimate (Est) of the mean population for counties in each subset based on 2016 ACS data

[2] STDEV = Standard Deviation of the Mean

[3] Proportion of population reporting each race alone or in combination with one or more other races

Table 2. County-Level Census Measures from 5-Year Estimates								
Characteristic	**Included Counties**		**Above Median RT**		**Below Median RT**		**Overall United States**	
Population	**Est[1]**	**STDEV[2]**	**Est**	**STDEV**	**Est**	**STDEV**	**Est**	**STDEV**
2016 ACS Estimate	711302	891526	805093	1156488	622102	517767	102930	331151
Race/Ethnicity[3]	**Mean%**	**STDEV**	**Mean%**	**STDEV**	**Mean%**	**STDEV**	**Mean%**	**STDEV**
White	68.9	16.1	64.0	18.1	73.6	12.2	83.4	16.8
Black	16.0	14.7	20.3	16.7	12.0	11.0	9.0	14.5
Asian	5.5	5.9	5.4	6.8	5.7	4.9	1.3	2.7
Hispanic	18.8	19.0	19.0	20.5	18.6	17.5	8.9	13.6
Place of Birth	**%**		**%**		**%**		**%**	
US Born	86.0	9.0	86.2	10.4	85.8	7.5	95.4	5.7
Non-US Born	14.0	9.0	13.8	10.4	14.2	7.5	4.6	5.7
Occupants Per Room	**%**		**%**		**%**		**%**	
More than 1	3.7	3.6	4.3	4.6	3.2	2.3	2.4	2.4
Employment	**%**		**%**		**%**		**%**	
Unemployed	4.8	1.3	5.1	1.4	4.6	1.1	4.0	1.7
Employed	59.0	6.2	57.6	6.4	60.3	5.6	54.5	8.4
Income	**USD ($)**		**USD ($)**		**USD ($)**		**USD ($)**	
Median Income	59008	16878	53642	13930	64111	17872	47973	12606
Health Insurance	**%**		**%**		**%**		**%**	
Insurance	87.9	5.2	86.6	5.4	89.1	4.8	87.7	5.3
No Insurance	12.1	5.2	13.4	5.4	10.9	4.8	12.3	5.3
Proportion of Families below poverty level	11.6	5.3	13.5	5.6	9.8	4.3	12.0	5.8
Educational Attainment	**%**		**%**		**%**		**%**	
Proportion Bachelors or Higher	32.3	11.0	29.9	10.5	34.5	11.1	20.8	9.1
Proportion High School Graduate	86.8	6.7	85.3	6.8	88.3	6.2	85.8	6.5

[1] Estimate (Est) of the mean population for counties in each subset based on 2016 ACS data

[2] STDEV = Standard Deviation of the Mean

[3] Proportion of population reporting each race alone or in combination with one or more other races

Table 3. Data Sources and Variable Names			
Variable Name	**Data Source**	**Name in Source**	**Notes**
Income	ACS_16_5YR_DP03	HC01_VC85	Median household income
Employed	ACS_16_5YR_DP03	HC03_VC06	Percent over 16yo in Labor Force, Employed
Unemployed	ACS_16_5YR_DP03	HC03_VC07	Percent over 16yo in Labor Force, Unemployed
Insured	ACS_16_5YR_DP03	HC03_VC131	Percent Civilian Population with Health Insurance
Uninsured	ACS_16_5YR_DP03	HC03_VC134	Percent Civilian Population without Health Insurance
Poverty	ACS_16_5YR_DP03	HC03_VC161	Percent of Families Below Poverty Line in the last 12 months
US-Born	ACS_16_5YR_DP02	HC03_VC131	Percent US Born
Non-US Born	ACS_16_5YR_DP02	HC03_VC136	Percent Non-US Born
White	ACS_16_5YR_DP05	HC03_VC49	Percent Reporting White Race
Black	ACS_16_5YR_DP05	HC03_VC50	Percent Reporting Black Race
Asian	ACS_16_5YR_DP05	HC03_VC56	Percent Reporting Asian Race
Hispanic	ACS_16_5YR_DP05	HC03_VC88	Percent Reporting Hispanic Ethnicity
High School Graduate or Higher	ACS_16_5YR_S1501	HC02_EST_VC17	Percent High School Graduate or Higher
Bachelor's Degree or Higher	ACS_16_5YR_S1501	HC02_EST_VC18	Percent Bachelor's Degree or Higher
Crowding	ACS_16_5YR_B25014	Sum HD01_VD05, 06,07,11,12,13	Percentage of households with 1.01 occupants or more per room

Table 4. County-Level Census Measures[1]								
Predictor	**Above Median RT, unadjusted univariate**			**Above Median RT, adjusted multivariate**			**Above Median RT, multivariate parsimonious**	
Race/Ethnicity[2]	**OR[3]**	**95% CI**	**P**	**aOR[3]**	**95% CI**	**P**	**aOR[3]**	**95% CI**
Hispanic	1	0.89-1.14	0.864	0.7	0.46-1.01	0.058	0.7	0.52-0.91
Black	1.5	1.28-2.89	<0.001	1.4	0.99-1.88	0.061	1.3	1.01-1.74
Asian	0.9	0.62-1.38	0.698	1.6	0.74-3.46	0.229	-	-
Place of Birth								
Non-US Born	1	0.73-1.24	0.707	1.1	0.59-2.21	0.687	-	-
SES								
More than 1 Occupant per Room	2.2	1.15-4.14	0.018	3.1	0.54-17.21	0.206	6.2	1.65-23.16
Unemployed	2.6	1.40-4.76	0.002	0.5	0.16-1.33	0.152	-	-
Proportion without Health Insurance	2.8	1.66-4.74	<0.001	1.7	0.65-4.42	0.3	-	-
Proportion Below Poverty Level	5	2.79-8.98	<0.001	5.5	1.74-17.29	0.004	4.4	1.96-10.01
Proportion Less Than Bachelor's	0.7	0.53-0.84	<0.001	0.8	0.54-1.25	0.369	-	-

[1] Characteristics from 2016 ACS 5-Year Estimates

[2] Proportion of population reporting each race alone or in combination with one or more other races

[3] Odds ratio represents a 10 percent change in each predictor except for unemployment and more than 1 occupant per room which represent 5 percent changes

References

1. Delogu G. The Biology of Mycobacterium Tuberculosis Infection. *MJHID.* 2013;5(1):e2013070.
2. Jong, B. Differences between TB cases infected with M. africanum, West-African type 2, relative to Euro-American M. tuberculosis- an update. *FEMS Immunol Med Microbiol*. 2010;58(1):102-105.
3. Gutierrez MC, Brisse S, Brosch R, Fabre M, Omais B, Marmiesse M, Supply P, Vincent V. Ancient origin and gene mosaicism of the progenitor of *Mycobacterium tuberculosis*. *PLoS Patholog*. 2005;1(1):e5.
4. Blaser MJ, Kirschner D. The equilibria that allow bacterial persistence in human hosts. *Nature*. 2007;449(7164):843-9.
5. Korch SB, Contreras H, Clark-Curtiss JE. Three Mycobacterium tuberculosis Rel toxin-antitoxin modules inhibit mycobacterial growth and are expressed in infected human macrophages. *J Bacteriol.* 2009;191(5):1618-30.
6. Frahm M, Goswami N. Discriminating between latent and active tuberculosis with multiple biomarker responses. *Tuberculosis (Edinb).* 2011;91(3):250-6.
7. Ai JW, Ruan QL, Liu QH, Zhang WH. Updates on the risk factors for latent tuberculosis reactivation and their managements. *Emerg Microbes Infect*. 2016;5:e10.
8. Sandgren A, Noordegraaf-Schouten MV, van Kessel F. Initiation and completion rates for latent tuberculosis infection treatment: a systematic review. *BMC Infect Dis*. 2016;16:204.
9. Tasilli A, Salomon JA, Trikalinos TA. Cost-effectiveness of Testing and Treatment for Latent Tuberculosis Infection in Residents Born Outside the United States with and Without Medical Comorbidities in a Simulation Model. *JAMA Intern Med.* 2017;177(12):1755-1764.
10. Munoz L, Stagg HR, Abubakar I. 2015. Diagnosis and Management of Latent Tuberculosis Infection. *Cold Spring Harb Perspect Med*. 2015;5(11):a017830.
11. Sia IG, Wieland ML. Current Concepts in the Management of Tuberculosis. *Mayo Clin Proc*. 2011;86(4):348-361.
12. Fox GJ, Barry SE, Britton WJ, Marks GB. Contact Investigation for Tuberculosis: a systematic review and meta-analysis. *Eur Respir J.* 2013;41(1):140-56.
13. Kelly AM, D'Agostino JF, Andrada LV, Liu J, Larson E. Delayed tuberculosis diagnosis and costs of contact investigations for hospital exposure: New York City, 2010-2014. *Am J Infect Control.* 2017;45(5):483-486.
14. Bloch AB. Screening for Tuberculosis and Tuberculosis infection in High-Risk Populations Recommendations of the Advisory Council for the Elimination of Tuberculosis. *MMWR Recomm Rep.* 1995;44(RR-11):19-34.
15. Mathema B, Andrews JR, Cohen T, Borgdorff MW, Behr M, Glynn JR, Rustomjee R, Silk BJ, Wood R. Drivers of Tuberculosis Transmission. *J Infect Dis.* 2017;216(suppl_6):S644-S653.
16. Centers for Disease Control and Prevention (CDC). *Reported Tuberculosis in the United States, 2017.* Atlanta, GA: US Department of Health and Human Service, CDC; 2018.

17. Stewart RJ, Tsang CA, Pratt RH, Price SF, Langer AJ. Tuberculosis – United States, 2017. *MMWR Morb Mortal Wkly Rep.* 2018;67:317-323
18. Centers for Disease Control and Prevention (CDC). Tuberculosis Genotyping in the United States, 2004-2010. Atlanta, GA: US Department of Health and Human Services, CDC, June 2012.
19. US Department of Health and Human Services, Centers for Disease Control and Prevention. Guide to the Application of Genotyping to Tuberculosis Prevention and Control. https://www.cdc.gov/tb/programs/genotyping/chap3/3_cdclab_2description.htm. Published September 1, 2012. Accessed April 1st, 2019.
20. Talrico S, Silk B. Whole-genome sequencing for investigation of recent TB transmission in the United States: Current uses and future plans [Presentation]. Presented at TB PEN Focal Point Call. February 7, 2018.
21. France, A. M., Grant, J., Kammerer, J. S., & Navin, T. R. (2015). A Field-Validated Approach Using Surveillance and Genotyping Data to Estimate Tuberculosis Attributable to Recent Transmission in the United States. American Journal of Epidemiology,182(9), 799-807. doi:10.1093/aje/kwv121
22. Yuen CM, Kammerer JS Marks K, et al. Recent Transmission of Tuberculosis – United States, 2011-2014. *PLoS One.* 2016;11(4):e0153728
23. Oren E, Masahiro N, Charles N, et al. Neighborhood Socioeconomic Position and Tuberculosis Transmission: A Retrospective Cohort Study. *BMC Infect Dis.* 2014;14:227.
24. Myers WP, Westenhouse JL, Flood J, et al. An Ecological Study of Tuberculosis Transmission in California. *Am J Public Health.* 2006;96(4):685-690.
25. Millet JP, Moreno A, Fina L, et al. Factors that Influence Current Tuberculosis Epidemiology. *Eur Spine J.* 2013;22(Suppl 4):539-548.
26. Olson NA, Davidow AL, Winston CA, et al. A National Study of Socioeconomic Status and Tuberculosis Rates by Country of Birth, United States, 1996-2005. *BMC Public Health*. 2012;12:365.
27. Moonan PK, Ghosh S, Oeltmann JE, et al. Using Genotyping and Geospatial Scanning to Estimate Recent Mycobacterium tuberculosis Transmission, United States. *Emerg Infect Dis.* 2012;18(3):458-465.