UNIVERSITY OF CALIFORNIA

Los Angeles

Tuberculosis Disease Incidence Estimation among Foreign-born Persons,
Los Angeles County 2005-2011

A dissertation submitted in partial satisfaction of the requirements
for the degree of Doctor of Philosophy in Epidemiology

by

Adam Readhead

2017

ABSTRACT OF THE DISSERTATION

Tuberculosis Disease Incidence Estimation among Foreign-born Persons,
Los Angeles County 2005-2011

by

Adam Readhead
Doctor of Philosophy in Epidemiology
University of California, Los Angeles 2017
Professor Roger Detels, Co-Chair
Professor Frank J. Sorvillo, Co-Chair

Tuberculosis is a global public health issue with more than 2 billion people infected worldwide. It is also a serious public health concern within the United States with 9,557 cases of active disease diagnosed in 2015 alone [1]. In the U.S., specific sub-groups, such as foreign-born persons, persons with diabetes or persons living with HIV or other immunocompromising conditions are known to be at higher risk of TB disease. Among foreign-born residents in the U.S., persons born in high-morbidity countries are known to be at even higher risk of developing the disease. Yet, TB disease incidence rates by country of birth are not reported at the local, state or national level despite these large, known differences in risk by country of birth. This is part due to the complications of using country-of-birth-specific population estimates and technical challenges of using standard regression analysis with a communicable disease. This thesis aims to call attention to this notable gap and, in part, to fill it.

Data on 5,447 diagnosed TB cases from the Los Angeles County Department of Public Health TB Control Program were combined with stratified population estimates available from the Public Use Microdata Survey to calculate the incidence rate of TB disease for the years 2005 through 2011, stratifying by country of birth and other demographic factors. Bayesian models were used to account for the uncertainty in the number of diagnoses and the population estimates. Extending these models into spatial analysis required the use of a hierarchical Bayesian model. Prediction models were constructed using bootstrap backward elimination and stochastic variable selection.

We estimated that the unadjusted incidence rate among persons born in the Philippines was 44.3 per 100,000 person-years and among persons born in Vietnam 38.7 per 100,000 person-years in comparison to 2.3 per 100,000 for U.S.-born persons. In spatial analysis, TB disease incidence was found to be spatially heterogeneous within Los Angeles County and remained so within high-risk countries of birth and when accounting for age, sex and years in residence. In prediction modeling, we found the addition of PUMA-level ecological variables did not improve the prediction of TB disease incidence beyond models using age, sex, country of birth and years in residence. With these three analytical approaches–non-spatial, spatial and prediction–we confirmed that TB disease incidence rates varied markedly by country of birth and showed that issues arising from the technical challenges of dependent outcomes, sparse data and uncertainty in population estimates can be ameliorated.

The dissertation of Adam Readhead is approved.

Karin Nielsen

Judith Silverstein Currier

Onyebuchi Aniweta Arah

Robert J. Kim-Farley

Frank J. Sorvillo, Committee Co-Chair

Roger Detels, Committee Co-Chair

University of California, Los Angeles

2017

To Heather, who never lets me give up on my dreams.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACS	American Community Survey
BCG	bacille Calmette-Guerin
CalREDIE	California Reportable Disease Information Exchange
CDC	Centers of Disease Control and Prevention
CDPH	California Department of Public Health
IGRA	Immunoglobulin Release Assay
IPUMS	Integrated Public Use Microdata Series
LAC DPH	Los Angeles County Department of Public Health
LTBI	Latent Tuberculosis Infection
NAAT	Nucleic Acid Amplification Test
OSHA	Occupational Safety and Health Administration
PUMA	Public Use Microdata Area
PUMS	Public Use Microdata Survey
RVCT	Reported Verified Case of Tuberculosis
TB	Tuberculosis
TBCB	Tuberculosis Control Branch
TBCP	Tuberculosis Control Program
TRIMS	Tuberculosis Reporting Information Management System
TIMS	Tuberculosis Information Management System
TST	Tuberculosis Skin Test

Acknowledgements

I would like to acknowledge my teachers, colleagues, family and friends who helped me complete this thesis. Thanks to Dr. Roger Detels, who gave me room to explore and guidance along the way; Dr. Sorvillo, for his unstinting support through the twists and turns of the project; Dr. Kim-Farley, who has been so generous with his time and support and would not let me drop my petition to transfer into the PhD program; Dr. Nielsen, for her thoughtful comments and encouragement; Dr. Currier, for her ability to ground the research in practice and Dr. Arah, for his commitment to tackling the difficult topics and not letting his students shy away from them.

Special thanks to my qualifying exam study group, Andrew, Aileen, Rufaidah, Laura and Amanda, for your support, good humor, and measured correction of erroneous ideas.

VITA

EDUCATION & PROFESSIONAL EXPERIENCE

1999	Bachelor of Arts, University of California at Berkeley
2002-2004	Research Assistant, HIV/AIDS Program Louisiana Office of Public Health
2004	Master of Public Health, Tulane University
2004-2007	Data Manager, Reproductive Health and HIV Research Unit (RHRU) Johannesburg, South Africa
2008-2011	Epidemiologist, HIV Surveillance and Field Services, New York City Department of Health and Mental Hygiene
2013	Epidemiology Analyst, TB Control Program Los Angeles Department of Public Health
2014-2016	Senior Research Scientist Spokane Regional Health District
2017-	Senior Epidemiologist, TB Control Branch California Department of Public Health

SELECTED PUBLICATIONS

Readhead AC, Gordon DE, Wang Z, Anderson BJ, Brousseau KS, Kouznetsova MA, Forgione LA, Smith LC, Torian LV. Transmitted antiretroviral drug resistance in New York State, 2006-2008: results from a new surveillance system. PLoS One. 2012;7(8):e40533. Epub 2012 Aug 6.

Tomines A, Readhead H, **Readhead A**, Teutsch S. Applications of Electronic Health Information in Public Health: Uses, Opportunities and Barriers. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2013;1(2). Article 5.

SELECTED PRESENTATIONS & POSTERS

Readhead A, Readhead H, Tomines A, Kim-Farley R. Assessment of a large local health department's need for electronic medical record data. APHA 2012, San Francisco.

Xia Q, **Readhead A**, Wiewel E, Torian L. Late stage HIV infection detection rate in New York City, 2004-2008. 19th International AIDS Conference 2012, Washington DC. Abstract no. TUPE227

Delany S, Moyes J, Pascoe S, **Readhead A**, Rees H. Sexually transmitted infections may play an important role in mediating the effect of microbicides in Phase III trials. Microbicides 2006, Cape Town.

Introduction

Tuberculosis is a global public health issue with more than 2 billion people infected, 10.4 million new diagnoses of the disease and 1.4 million deaths in 2015 [2]. It has surpassed HIV as the number one infectious disease killer [3]. In the first half of the 20th century, the U.S. made substantial progress towards control of the tuberculosis epidemic domestically capitalizing on advances in bacteriology and improved living conditions together with using an aggressive public health campaign to humble the dreaded "white plague" that contributed to an estimated "40% of working-class deaths in cities" in the 19th century [4] . But TB eradication in the U.S. has been elusive [5]:

> …[F]unding for TB research and treatment dropped from US$ 40 million a year in the late 1960s to only US$ 283 000 in 1980, while in 1989 the US Department of Health and Human Services was so confident TB was finally on the run that it predicted TB would be more or less eradicated from the country by 2010.

While TB disease incidence is low among U.S.-born persons in general, TB is a significant and seemingly intractable issue among foreign-born residents of the U.S. TB burden remains a concern in states like New York, California, Texas and Florida because of reactivation of latent TB among long-term foreign-born residents and continued immigration from medium and high TB prevalence countries. Tuberculosis in Los Angeles County is of special concern because more than a one-third of County residents - over three and half million people - are foreign-born [6].

Background

Background on Mycobacterium tuberculosis

Tuberculosis (TB) is caused by *Mycobacterium tuberculosis*, a slow-growing, acid-fast, rod-shaped mycobacterium that most commonly infects the lungs but can also infect other organs including the lymph nodes, the brain and the blood system [7]. Other species of mycobacterium, such as *Mycobacterium bovis* and *Mycobacterium avium*, can produce tuberculosis-like disease in humans. Tuberculosis is transmitted person-to-person when tubercle bacilli in a droplet nuclei are expelled from the infected individual's lung, through activities such as coughing, sneezing, speaking and singing, and subsequently inhaled by another individual [8]. Persons in close contact with an infected individual are at risk of becoming infected.

Most persons infected with TB develop a latent tuberculosis infection (LTBI) in which the *M. tuberculosis* bacteria is resident in the body but contained by the immune system in a granuloma. Persons with LTBI are not infectious to others. In part because the organism is slow-growing, an individual can remain infected for years without manifest disease. TB may re-emerge if an individual's immune system is weakened either through natural senescence, immune-compromising diseases such as HIV or diabetes mellitus, or immunosuppressive drugs used in cancer chemotherapy, organ transplant or patients with rheumatoid arthritis. In the 1960s, TB researchers used observational studies to estimate the lifetime risk of reactivation of latent infections at 5-10% [9]. This estimate has become the benchmark for reactivation risk, but more recent studies suggest a lifetime risk of approximately 1% [10, 11]. This is an area of active research in the field.

Other persons infected with TB develop active disease. The highest risk of progression to active disease is within the first two years of the infection [12]. Approximately 80% of all cases develop active disease within two years of being infected. Those with comorbidities such as HIV and diabetes are more likely to progress to active disease.

Epidemiology of Tuberculosis

More than one-third of the world's population, 2 billion people, is infected with *Mycobacterium tuberculosis*. In 2015, there were 10.4 million new cases of tuberculosis disease worldwide [2]. The tuberculosis disease incidence rates vary greatly by country, ranging from <1 case per 100,000 person-years to over 800 cases per 100,000 person-years in South Africa [13]. At the turn of the 20th century, tuberculosis was a leading cause death in the United States, but improved living conditions, pasteurization of milk, as well as the discovery of potent treatments and an aggressive control campaign contributed to a steep decline in domestic TB cases between 1900 and 1960. However, TB remains a significant health burden among foreign-born, homeless, incarcerated and/or HIV-positive populations. Nationally, there were 9,557 new TB cases in 2015 with an incidence rate of 3 cases per 100,000 [14]. This incidence rate is one of the lowest in the world and is on par with the case rates in Iceland and Israel [13]. Most of the cases in the U.S. were foreign born and, of foreign-born cases, most were from Mexico, Philippines, India, Vietnam or China [14]. TB disease incidence rate among the foreign born was 15.1 per 100,000 compared to 1.2 per 100,000 among the U.S. born [14]. These countries have higher case rates than the U.S.; the rates per 100,000 person-years were 21 in Mexico, 322 in the Philippines, 217 in India, 137 in Vietnam and 67 in China [13]. California, Texas, New York and Florida accounted for half of all reported cases because of the large number of foreign-born residents in these jurisdictions [14].

This research focused on Los Angeles County from 2005-2011. In 2011, foreign-born persons accounted for 66% of reported verified cases of tuberculosis nationally and 78% of cases in Los Angeles County [15-17]. In 2012, foreign-born persons had a diagnosis case rate of 15.9 per 100,000, in contrast to 1.4 per 100,000 for U.S.-born persons highlighting tuberculosis as an important factor in health disparity [15].

Legal and Regulatory Precedence related to Tuberculosis Control

TB control is one of the oldest public health efforts in the US and has a long legal history [18]. All states legally required reporting of TB as of 1901 and national surveillance of the disease began in 1953 [19]. TB-related regulation is found throughout the California Health and Safety Code [20, 21]. Recent legal changes in California have been driven by concerns about occupational safety. In 1997, the Occupational Safety and Health Administration (OSHA) released a Proposed Rule on Occupational Exposure to Tuberculosis, but a federal standard for TB control in the workplace was never established [22]. In 2001, OSHA withdrew the proposed rule, claiming that the risk of infection had been overestimated and the continued decline of TB nationwide greatly reduced the risk overall. California OSHA decided to follow through with the standard and added TB language to the OSHA-approved state plan [23]. California OSHA used the TB standard and widened the scope to create the Aerosol Transmissible Disease Standard which is listed in the California Code of Regulations, Title 8, Section 5199 Aerosol Transmissible Diseases [24].

Priorities for TB control

The Centers for Disease Control and Prevention (CDC), American Thoracic Association and the Society for Infectious Disease established the following priorities for TB control [25]:

1. Early and accurate detection, diagnosis, and reporting of TB cases leading to initiation and completion of treatment;
2. Identification of contacts of patients with infectious TB and treatment of those at risk with an effective drug regimen;
3. Identification of other persons with latent TB infection at risk for progression to TB disease and treatment of those persons with an effective drug regimen;
4. Identification of settings in which a high risk exists for transmission of *Mycobacterium tuberculosis* and application of effective infection-control measures.

Methods

Data Sources

Data Collection, Processing and Management

Aggregate and case-specific data on foreign-born reported verified cases of tuberculosis (RVCTs) diagnosed from 2005 to 2011 were drawn from the tuberculosis surveillance system at the Los Angeles County Department of Public Health, Tuberculosis Control Program (TBCP). Medical providers are required by law to report tuberculosis diagnoses to the TBCP [14, 21].

Basic demographic and risk data are reported to TBCP by the diagnosing provider using a faxed reporting form or by telephone interview. In some cases, additional data may be collected through telephone interview with the diagnosing provider at the time of report especially if required data is missing. Case-specific data includes age at diagnosis, sex, address at diagnosis, length of residence in the U.S., year of diagnosis, smear status, disease site and resistance profile. Program staff verify the case through laboratory testing or other methods. Note that in the case verification process cases that do not reside in Los Angeles County are re-assigned to their county of residence through a routine reconciliation process conducted at regular intervals. Upon receipt of a putative new case, the case is checked against existing cases to ensure that it has not been previously reported. If no matching case is found among the existing cases, an investigation is initiated. TBCP staff inform the assigned community health services (CHS) nurse that a case needs verification. The nurse will travel to the case location and verify the case. A contact investigation is initiated if necessary. In the course of the contact investigation, the public health nurse or public health investigator will interview the patient and may confirm or correct previously-reported data. Data from confirmed and suspected cases of TB are entered into the Tuberculosis Reporting Information Management System (TRIMS). Cases that are verified per CDC’s RVCT criteria are reported to the California Department of Public Health, TB Control Branch, who aggregate and transmit redacted data to the Centers for Disease Control and Prevention for national surveillance purposes (see Appendix 1 for CSTE 2009 case definition).

For this analysis, cases are defined as only those reported verified cases of tuberculosis (RVCT) that were reported to the California Department of Public Health, TB Control Branch as of July 2014. The case data reported to the state was used for address data because the address from TRIMS may not be the address at diagnosis. The address data in TRIMS is the current address of the case; if a case's address changes, the new address overwrites the previous address. There were two data systems for reported cases at the state level during the study period. The first was the TB Information Management System (TIMS) which contains data on cases from 1997 through 2007. The second is the California Reportable Disease Information Exchange (CalREDIE) which contains data on cases from 2008 to the present. Identifiers for these cases, together with reported address data, will be extracted from these databases and matched to TRIMS. All data with the exception of case identifiers and case address were drawn from TRIMS.

Address at diagnosis data was cleaned and geo-coded using a written protocol and the Los Angeles County geo-coding service. The geo-coded address was spatially joined with the Public Use Microdata Area (PUMA) boundaries for both the 2000 and the 2010 U.S. Census.

Data used to create population estimates of country of birth specific denominators were drawn from the U.S. Census's American Community Survey via the Minnesota Population Center at the University of Minnesota which maintains curated copies of the American Community Survey Public Use Microdata Survey (ACS PUMS) data in their Integrated Public Use Microdata Series (IPUMS) [26]. The American Community Survey is telephone-based survey which interviews a 1% sample of U.S. households each year. Some individual survey data are released for public use through the Public Use Microdata Survey (PUMS). These data are manipulated to prevent re-identification of individual respondents while preserving the original distributions of measurements. In this analysis, the PUMS data was be used in conjunction with the Public Use Microdata Areas (PUMA). Each PUMA is derived by grouping census tracts and is designed to encompass approximately 100,000 persons. The boundaries files for the LA Country PUMAs were sourced from the U.S. Census.

Study Setting

The study was set in Los Angeles County, a large, diverse county with a sizeable foreign-born population. The population of Los Angeles County was estimated to be 10,137,915 persons in 2016 [27]. By population, the county is larger than 41 states [28]. The county covers an area of 4,057 square miles and includes 88 cities [27] [29]. The population size, the large geographic area and the complex political and regulatory environment of the County pose significant challenges for local public health agencies in terms of service provision, program implementation and policy change.

The Los Angeles-Anaheim-Long Beach urbanized area was the densest urban area in the nation according to the 2010 Census [30]. The population density of the county is 2,419 person per square mile, but ranges from approximately 20 persons per square mile in Vernon to 42,611 in Koreatown [31]. The Westlake neighborhood (directly east of Korea town) is the second most dense neighborhood with 38,214 persons per square mile, roughly comparable in density to the Brooklyn, New York which has 37,660 persons per square mile [32] [33].

Of the more than 10 million persons residing in the County, 35.3%, or approximately 3,536,025 persons, are estimated to be foreign-born.[6] Of foreign-born persons residing in LA County between 2008 and 2010, 77% had immigrated to the area since 1980; 92% had immigrated since 1970.[6] The number of foreign-born persons in Los Angeles County has more than doubled between 2000 and 1980.[6] Of foreign-born persons residing in the County in 2012, 67% were from six countries: Mexico (39%), El Salvador (7%), the Philippines (7%), Guatemala (5%) and South Korea (5%) and China excluding Hong Kong and Taiwan (4%) [34]."Linguistic isolation – the proportion of immigrant-headed households in which no person over 13 speaks English only, or very well – is relatively high at 34%."[6]

Foreign-born persons are not uniformly distributed throughout the county. A number of foreign-born communities exist including Artesia, Boyle Heights, East Los Angeles, Fairfax, Koreatown, Japantown, Palos Verdes, San Gabriel, Monterey Park, Alhambra, Glendale, Burbank and Westwood.

The Los Angeles County Department of Public Health is a large local health department with more than 4,000 employees dedicated to the protection and improvement of the health of residents of Los Angeles County. The stated mission of the department is "to protect health, prevent disease, and promote health and well-being." [35] Within the Los Angeles Country Department of Public Health, the Division of Communicable Disease Control and Prevention is responsible for the surveillance and control of communicable diseases except for sexually transmitted diseases including HIV/AIDS which is housed in a separate division. Within the division, the Tuberculosis Control Program (TBCP) is responsible for the routine surveillance of tuberculosis as well as planning special TB investigations and providing medical consultations to providers treating TB cases. The Community Health Services division is responsible for routine tuberculosis contact investigations in coordination with TBCP as well as providing clinical services from its 14 clinics.

Study Design

The study used a serial cross-sectional design with a study period of 2005 through 2011. Data was drawn from routine tuberculosis surveillance conducted by Los Angeles County Department of Public Health and from the U.S. Census Bureau's American Community Survey (ACS). Individual-level data is available for persons with reported, verified cases of tuberculosis (RVCT). Detailed description of the data collection and data analysis is the following chapters. Briefly, we estimate TB disease incidence among foreign-born persons stratifying by key factors of interest including: age at diagnosis, sex, length of residence in the U.S. and PUMA.

TB Disease Incidence by Country of Birth, Los Angeles County 2005-2001

Abstract

Introduction

Among U.S. residents, tuberculosis disease incidence rate is higher among foreign-born persons than U.S.-born persons and varies substantially by country of birth, yet local health departments seldom report TB disease incidence rates by country of birth. With more than 3.5 million foreign-born residents, Los Angeles County has the largest number of foreign-born persons of any U.S. county and contributes roughly 7% of all cases of TB disease nationally. Further description of local TB burden including incidence rates by country of birth would aid continued public health response.

Methods

Data on 5,447 diagnosed TB cases from the Los Angeles County Department of Public Health TB Control Program were combined with stratified population estimates available from the Public Use Microdata Survey to calculate incidence rate of TB disease for the years 2005 through 2011. Unadjusted incidence rates were calculated by country of birth and other demographic factors. To mitigate issues stemming from correlated outcomes, incidence rates were modelled using a negative binomial regression. Bayesian models were used to account for the uncertainty in the number of diagnoses and the population estimates.

Results

Unadjusted incidence rates among several foreign-born populations were notably higher than among U.S.-born persons; the unadjusted incidence rate was 44.3 per 100,000 person-years among persons born in the Philippines and 38.7 per 100,000 person-years among persons born in Vietnam in comparison to 2.3 per 100,000 for U.S.-born persons. The largest absolute number of cases of TB disease was among Mexican-born persons (n=1,234); the unadjusted incidence rate in this group was 12.4 per 100,000 person-years. Accounting for age, gender, length of residence and year of diagnosis, persons born in Vietnam were 4.5 (95% CI: 3.8 – 5.3) times as likely to have been diagnosed with active TB disease than persons born in countries other than the eight countries reporting highest rates of TB disease. In contrast, persons born in Mexico were 1.67 (95% CI: 1.47 – 1.89) times as likely as other foreign-born persons to be diagnosed with active TB. Bayesian models showed similar results.

Conclusion

This study confirms that incidence of TB disease varied markedly by country of birth in Los Angeles County. Even accounting for differences in age, gender, years in residence distributions, persons from the Philippines, Vietnam and other countries were at greater risk of TB disease than persons from other foreign countries. We have also demonstrated that the disparity in risk by country of birth can be readily estimated using available data and that complex adjustment of denominator error using Bayesian techniques has limited utility.

Introduction

Tuberculosis (TB) is a global public health issue with more than one-third of the world infected with the mycobacterium. In 2015, there were 10.4 million new diagnoses of the disease [2] and it surpassed HIV as the number one infectious disease killer, implicated in 1.4 million deaths [2]. In low incidence countries, such as the U.S., the majority of TB diagnoses occur among foreign-born persons [14, 36, 37]. Los Angeles County is home to 3.5 million foreign-born persons, the largest concentration of foreign-born persons in the U.S. within a single county [38]. Earlier studies have shown substantial disparity in incidence rates of TB disease by country of birth both in Los Angeles County and nationally, yet state and local health departments do not report incidence rates by country of birth [39, 40]. The California TB Elimination plan identifies country of birth as an important risk factor and calls for additional data collection and analysis by country of birth. Here we describe the risk of TB disease in Los Angeles County by country of birth and demonstrate incidence rate estimation using available data and a range of models from simple to complex.

Methods

Description of data sources

Data on 5,447 reported verified cases of tuberculosis (RVCTs) diagnosed from 2005 through 2011 were drawn from the tuberculosis surveillance system of the Los Angeles County

Department of Public Health, Tuberculosis Control Program (TBCP). Tuberculosis surveillance has been described elsewhere [39, 40]. Briefly, medical providers are required by California State law to report persons with suspected or confirmed active TB disease and provide basic demographic and clinical information for these patients [20, 21, 41]. TBCP staff verify the diagnosis by reviewing laboratory, microbiology, and radiographic results or other clinical information, and may request additional diagnostics. Cases that meet the criteria for a report of a verified case of tuberculosis (RVCT) are reported to the California Department of Public Health, TB Control Branch, who in turn report the case to the Centers for Disease Control and Prevention [41].

Data used to create stratified population estimates were drawn from the U.S. Census Public Use Microdata Survey via the Minnesota Population Center at the University of Minnesota which maintains curated copies in their Integrated Public Use Microdata Series (IPUMS) [26, 42]. Population estimates by country of birth can also be obtained through American Factfinder (see Appendix 2 for additional information). The American Community Survey (ACS) is telephone-based survey which interviews a 1% sample of U.S. households each year [43]. Detailed data are released for public use through the Public Use Microdata Survey (PUMS), though these data are manipulated to prevent re-identification of individual respondents while preserving the original distributions of measurements. Population estimates were calculated using weighted frequencies. Confidence intervals for these estimates were calculated by standard and robust methods and replicate weights available with PUMS data [44, 45].

Beginning in 2006, the ACS sampling frame included both institutional and non-institutional group quarters. Before 2006, group quarters were not included in the ACS sampling frame. Cases residing in correctional facilities or long-term care were excluded if they were diagnosed in 2005 and included if they were diagnosed between 2006 and 2011. Cases that were homeless were excluded except for cases diagnosed between 2006 and 2011 and living in a homeless shelter.

We excluded a total of 494 (9%) cases: 26 cases with missing or unknown country of birth, 20 cases that were in correctional facilities (or missing this variable) and diagnosed in 2005, 16 cases that were in long-term care (or missing this variable), 277 cases that were homeless (or missing this variable) and not housed in group quarters, 13 cases diagnosed in 2005 that were homeless and living in group quarters, 14 administrative cases, 112 cases that were foreign-born but missing years in residence, 15 cases with countries of birth not included in the PUMS data and 1 case that could not be assigned to a PUMA. A total of 4,953 cases were available for analysis. Cases residing in Long Beach or Pasadena were not included as those cases were reported to the Pasadena City Health Department and the Long Beach Health and Human Services, respectively.

Data processing and definitions

Residential addresses at time of diagnosis were geocoded following a written protocol and using ArcGIS and the LA County-approved locator. Strata definitions for age, country of birth and length of residence were harmonized between TB surveillance data and PUMS population data. Data from South Korea and North Korea were combined into a single category because the American Community Survey did not separately enumerate North Koreans and South Koreans in the study period. Changes in country name during the study period were taken into account. Where individual country data were too sparse to produce a reliable estimate, data from multiple countries were aggregated to give a regional estimate, as was the case for several countries on the African continent. Isoniazid mono-resistance was defined as resistance to isoniazid only; multi-drug resistance (MDR) was defined as resistance to isoniazid and rifampicin with or without resistance to addition TB medications, and extreme drug resistance (XDR) was defined as meeting MDR criteria "plus any fluoroquinolone and at least one of three injectable second-line drugs" [46, 47]. Per the CDC, culture positive was defined as a positive culture from sputum or direct sputum collected within 15 calendar days of the start of treatment (if treatment was reported) or within 15 calendar days of diagnosis (if treatment was not reported) [41].

Analysis

We used serial cross-sectional design with a study period of 2005 through 2011. Unadjusted incidence rates were stratified by country of birth and other demographic factors and are presented without confidence intervals because data were extra-dispersed and therefore standard confidence intervals would likely understate the variability of the estimate. Relative standard error (RSE) was calculated as $\frac{1}{\sqrt{n}}$ where n is the number of cases [48]. Covariates for the generalized linear models (GLMs) were chosen *a priori* based on evidence in previous literature and on availability in both local TB surveillance and PUMS data [39, 40, 49]. Fitting a naïve Poisson GLM confirmed substantial residual over-dispersion. A Poisson model with country of birth and offset alone had a dispersion statistic of 10.5; in contrast, a Poisson model with all available covariates had a dispersion statistic of 1.5. We then fitted a negative binomial model, a common model for over-dispersed data [50]. To address under-coverage of confidence intervals due to residual extra-dispersion, robust confidence interval calculations were calculated per Hilbe [50].

To account for uncertainty in the population estimates, we adopted a Bayesian framework and introduced probability distributions or priors for these estimates. First, we built Bayesian analogues to the Poisson and negative binomial GLMs described above. Bayesian models were fitted using R, OpenBUGS, and nimble [51-53]. Standard BUGS coding was used (see Appendix 2). Priors for the intercept and all covariate coefficients were defined to be $N(0,1000)$. All covariates were categorical and had corner constraints on the reference category. For the negative binomial Bayesian model, the prior for r was $G(1,10)$. Two MCMC chains were run for 100,000 iterations each. Reasonable mixing and stability were achieved.

Second, we introduced informative priors on the population estimates. Theses priors were constructed to match the standard errors of these estimates calculated using replicate weights. The population estimate priors were truncated to [1, ∞] to ensure that the estimate was positive and non-zero. This resulted in a distortion of the prior distribution for some estimates.

The reference category for nativity was “other foreign-born countries” which was comprised of all available foreign countries except those with the highest absolute number of cases: the Philippines, Vietnam, India, China, Korea, Guatemala, Mexico and El Salvador. In keeping with guidelines adopted by the California Department of Public Health, TB Control Branch publications, tables were limited to denominator cell sizes of five or more [54].

Statistical analysis and data management were done using R version 3.4, R Studio version 1.0.143 and variety of packages [51, 55-65]. Bayesian models were run in OpenBUGS version 3.2.2 rev 1012 [52]. This study was deemed exempt by the Los Angeles County Department of Public Health Institutional Review Board.

Results

Unadjusted Analysis

The unadjusted TB disease incidence rate among foreign-born persons was 15.8 per 100,000 person years in the study period in comparison to 2.3 per 100,000 person years for U.S.-born persons (Table 1). In contrast to TB disease incidence rates among U.S.-born persons, the TB incidence rates were notably higher among persons born in several countries including Burma, Indonesia, Afghanistan, the Philippines, Vietnam, India and China and in Central, East and West Africa. Diagnoses of active TB among foreign-born persons accounted for approximately 80% of all cases in the study period. Of diagnoses among foreign-born persons, persons born in 8 countries accounted for more than 83% of cases in this time period (Table 1). Among these 8 countries, TB incidence rates ranged from 44.3 per 100,000 among Philippines-born persons to

9.4 per 100,000 among El Salvador-born persons. A more comprehensive list of countries of birth and associated incidence rates appears in table 2. TB disease incidence rates were very high among persons born in Central Africa and Burma, 168.5 per 100,000 and 78.9 per 100,000 respectively. Persons born in East Africa, Indonesia and Afghanistan also had high incidence rates at 47.6 per 100,000, 47.4 per 100,000 and 46.7 per 100,000, respectively. Note that several estimates, including estimates for Central Africa and Afghanistan, had a relative standard error of more than 30%, a common benchmark at which the reliability of an estimate is questioned [66, 67].

The proportion of culture positive cases that were isoniazid mono-resistant was notably higher among persons born in the Philippines, Vietnam or India in comparison to those born in Honduras, El Salvador or Mexico (Table 2). The proportion of isoniazid resistance ranged from 18% to 20% for active TB cases born in the Philippines, Vietnam or India. In contrast, the proportion of isoniazid resistance ranged from 3% to 8% for TB cases born in Honduras, El Salvador or Mexico. Similar patterns were observed for multi-drug resistance, though the number of cases was low: 0.7% of cases among persons born in Mexico were MDR, whereas 4.4% of cases among persons born in Korea were MDR. There were no cases with extreme drug resistance in the study period.

Among foreign-born persons, TB disease incidence rates were higher among older persons, men, and those residing in the U.S. for less than 2 years (Table 3). The unadjusted incidence rate for foreign-born persons 80 years old and older was 49.6 per 100,000 in contrast to an unadjusted incidence rate of 6.6 per 100,000 for persons under 20 years old. Foreign-born persons residing in the United States for less than two years had an unadjusted incidence rate of 69.8 per 100,000; those residing in the U.S. for two to four years had an incidence rate of 26.0 per 100,000. Of all diagnoses in the study period, 59% were among persons who had resided in the U.S. for 10 or more years. Incidence rates declined for both foreign-born persons and U.S.-born persons during the study period from 18.1 per 100,000 in 2005 to 14.0 per 100,000 in 2011.

Adjusted Analysis

We fit two generalized linear models to the data acknowledging that data with correlated outcomes, such as TB, are commonly over-dispersed. Estimates from the two models were similar. Dispersion statistics for both models were close to 1, though the negative binomial model was a better fit for the data (Table 4).

Accounting for other factors associated with TB incidence such as age, gender, length of residence and year of diagnosis, persons born in Vietnam were 4.5 (95% robust confidence interval 3.8 – 5.3) times as likely to have been diagnosed with active TB than persons born in countries outside of the TB top eight (table 4). In contrast, persons born in Mexico were 1.7 (1.5 – 1.9) times as likely as persons born in countries outside of the TB top eight (table 4) to be diagnosed with active TB.

Estimates calculated using Bayesian Poisson and negative binomial models were on par with the generalized linear models (Table 5). There were negligible differences between the Bayesian models based on Poisson and negative binomial distributions. Estimates using the Bayesian negative binomial model with informative priors for the population estimates were slightly decreased in comparison to the same model without these priors, likely due to the truncated distribution of the population estimate priors.

Discussion

In Los Angeles County, the burden of TB disease among foreign-born persons was much higher than among U.S.-born persons. In the study period, TB disease incidence was approximately 8 times more likely among foreign-born persons than among U.S.-born persons and 80% of reported cases of active TB were foreign-born. Similar results have been observed nationally and in other major metropoles [36, 39, 40]. Furthermore, incidence varied widely by country of birth. Even when accounting for other contributing factors including age, gender, and length of residence in the U.S., persons from the Philippines and Vietnam were approximately 4 times more likely than other foreign-born persons to have been diagnosed with TB. In contrast, persons born in India, China, Korea, Guatemala and Mexico were roughly twice as likely as other foreign-born persons to have been diagnosed with TB disease in the study period. This information can help guide local prevention efforts which are currently being planned as part of pilot latent TB testing programs.

Incidence rates declined for both foreign-born persons and U.S.-born persons during the study period from 18.1 per 100,000 in 2005 to 14.0 per 100,000 in 2011. There was a notable decline in TB disease incidence over the study period; analogous declines have been seen both nationally and in other cities though multiple contributing factors have been cited [36, 49]. Furthermore, we observed higher incidence rates among males, older adults and, most notably, those with fewer years in residence. This was consistent with several previous reports of TB incidence [40, 68]. The effect of age on incidence of TB disease has been unclear with numerous studies showing increased disease incidence among older persons, and other studies showing no increase with age among when restricted to certain diagnostic subgroups [69]. It is important to note that persons under 20 years of age tend to progress to disease more rapidly after infection, though there is no evidence of that here. While the highest incidence by years in residence was among those with less than 2 years in residence, the majority (59%) of foreign-born diagnoses had resided in the U.S. for 10 or more years.

The results also served to underscore the importance of testing for latent TB infection among foreign-born persons. The California Department of Public Health TB Control Branch has issued a tuberculosis risk assessment which recommends testing for foreign-born persons from countries "with an elevated TB rate", along with those who have or plan to have immunosuppression and those who are close contacts of an infectious case [70]. For medical providers, especially those serving foreign-born persons in Los Angeles County, detailed information on risk by country of birth, such as is provided in this analysis, may be helpful in a patent's overall risk assessment.

While based on mature surveillance and survey data, this study has several limitations. Despite adjusting for several influential factors, the data remained over-dispersed, which could result in the under-coverage of confidence intervals. Moreover, spatial effects and disease transmission were not taken into consideration which may further undermine these estimates. However, an estimated 85% of cases in California are due to reactivation and only 15% are due to recent transmission [71]. Thus, correlated outcomes typical of communicable diseases are less of a concern here. Biases due to misclassification and incomplete adjustment were not addressed. Case ascertainment for this surveillance system is unknown although we assume that it is high, similar to other TB surveillance systems [72-74]. Higher TB disease incidence among recently-immigrated foreign born could be in part the result of increased ascertainment in this group. Truncation of the population estimate prior to a positive, non-zero interval inflated some population estimates, but did so minimally. The presented models were constructed to be explanatory not predictive and, as such, are not the preferred set of models to use for planning disease interventions. In addition, foreign-born persons may be more likely to be screened, though certain foreign-born populations are less likely to have access to or use healthcare. Furthermore, some of the cases include in this analysis did not have culture positive results.

Conclusion

This study confirms that TB disease incidence varied markedly by country of birth in Los Angeles County. Even accounting for differences in age, gender, years in residence distributions, persons from the Philippines, Vietnam and other countries have much higher rates of reported TB disease than other foreign countries. Furthermore, TB disease incidence rates varied markedly by years in residence in Los Angeles County. Even accounting for other factors, persons with less than 2 years in residence presented with much higher incidence of TB disease than those in residence for 2 or more years. With this study, we established the relative strength of the key factors associated with TB diagnosis among the foreign-born and prepared the way for a model to predict future TB burden within this population. In addition, we have demonstrated that simple incidence rates by country of birth can be calculated with readily-accessible population estimates and that more complex adjusted estimates can be achieved the use of negative binomial models and Bayesian techniques. This analysis helped better describe the local TB burden in Los Angeles County and can be used to inform a continued public health response.

Table 1: TB Incidence Rates by Selected Country of Birth, Los Angeles County 2005*-2011

Country of Birth	Diagnoses	Person-Years	Proportion of Total Diagnoses	Unadjusted Incidence Rate per 100,000	95% Confidence Interval
Foreign country	3,946	25,037,400	80%	15.8	(15.3 - 16.3)
United States	1,008	43,987,963	20%	2.3	(2.2 - 2.4)
Philippines	742	1,674,344	15%	44.3	(41.1 - 47.5)
Vietnam	265	685,320	5%	38.7	(34.0 - 43.3)
India	98	331,396	2%	29.6	(23.7 - 35.4)
China	261	942,822	5%	27.7	(24.3 - 31.0)
Korea	249	1,123,255	5%	22.2	(19.4 - 24.9)
Guatemala	212	1,207,414	4%	17.6	(15.2 - 19.9)
Mexico	1,271	10,212,974	26%	12.4	(11.8 - 13.1)
Other foreign country	675	7,023,036	14%	9.6	(8.9 - 10.3)
El Salvador	173	1,836,839	3%	9.4	(8.0 - 10.8)
United States	1,008	43,987,963	20%	2.3	(2.2 - 2.4)

Source: Los Angeles County Department of Public Health, TB Control Program & Public Use Microdata Survey via IPUMS.

**For diagnosis year 2005, excluded cases indicated to be homeless, incarcerated or in long-term care facilities because ACS 2005 excluded these populations from group quarters.*

Table 2: TB Incidence Rates by Selected Country of Birth (Cases in Period >= 5), Los Angeles County 2005*-2011

Country of Birth	Diagnoses	Person-Years	Unadjusted Incidence Rate per 100,000	95% Confidence Interval	Proportion of Diagnoses	Culture Positive	Isoniazid Resistant		Multidrug Resistant	
							N	%	N	%
Central Africa	8	4,748	168.5	(51.7 - 285.3)	0.2	*	*	*	*	*
Burma (Myanmar)	34	43,072	78.9	(52.4 - 105.5)	0.7	22	*	*	*	*
East Africa	43	90,418	47.6	(33.3 - 61.8)	0.9	27	*	*	*	*
Indonesia	49	103,480	47.4	(34.1 - 60.6)	1.0	34	*	*	*	*
Afghanistan	9	19,253	46.7	(16.2 - 77.3)	0.2	*	*	*	*	*
Philippines	742	1,674,344	44.3	(41.1 - 47.5)	15.0	447	90	20.1	13	2.9
Vietnam	265	685,320	38.7	(34.0 - 43.3)	5.3	168	32	19.0	*	*
West Africa	26	87,117	29.8	(18.4 - 41.3)	0.5	9	*	*	*	*
India	98	331,396	29.6	(23.7 - 35.4)	2.0	32	6	18.8	*	*
China	261	942,822	27.7	(24.3 - 31.0)	5.3	177	11	6.2	6	3.4
Cambodia (Kampuchea)	43	173,988	24.7	(17.3 - 32.1)	0.9	25	*	*	*	*
Pakistan	12	50,634	23.7	(10.3 - 37.1)	0.2	5	*	*	*	*
Peru	46	195,596	23.5	(16.7 - 30.3)	0.9	30	*	*	*	*
Korea	249	1,123,255	22.2	(19.4 - 24.9)	5.0	180	22	12.2	8	4.4
Honduras	51	234,421	21.8	(15.8 - 27.7)	1.0	39	*	*	*	*
Thailand	29	155,805	18.6	(11.8 - 25.4)	0.6	18	*	*	*	*
Guatemala	212	1,207,414	17.6	(15.2 - 19.9)	4.3	145	12	8.3	*	*
Mexico	1,271	10,212,974	12.4	(11.8 - 13.1)	25.7	750	63	8.4	5	0.7
Belize/British Honduras	11	94,077	11.7	(4.8 - 18.6)	0.2	5	*	*	*	*
Taiwan	55	476,938	11.5	(8.5 - 14.6)	1.1	40	*	*	*	*
Colombia	12	111,087	10.8	(4.7 - 16.9)	0.2	6	*	*	*	*
Ecuador	10	94,998	10.5	(4.0 - 17.1)	0.2	*	*	*	*	*
El Salvador	173	1,836,839	9.4	(8.0 - 10.8)	3.5	101	*	*	*	*
Armenia	40	429,402	9.3	(6.4 - 12.2)	0.8	30	*	*	*	*
Hong Kong	18	193,797	9.3	(5.0 - 13.6)	0.4	10	*	*	*	*
Nicaragua	15	210,585	7.1	(3.5 - 10.7)	0.3	11	*	*	*	*

Iran	49	746,784	6.6	(4.7 - 8.4)	1.0	31	*	*	*	*
Japan	21	317,750	6.6	(3.8 - 9.4)	0.4	15	*	*	*	*
Russia	9	157,397	5.7	(2.0 - 9.5)	0.2	*	*	*	*	*
Argentina	6	107,344	5.6	(1.1 - 10.1)	0.1	*	*	*	*	*
North Africa	6	124,016	4.8	(1.0 - 8.7)	0.1	*	*	*	*	*
United States Outlying Areas	6	138,068	4.3	(0.9 - 7.8)	0.1	*	*	*	*	*
Cuba	5	148,281	3.4	(0.4 - 6.3)	0.1	*	*	*	*	*
Germany	5	176,771	2.8	(0.3 - 5.3)	0.1	*	*	*	*	*
United States	1,002	43,849,895	2.3	(2.1 - 2.4)	20.2	464	22	4.7	*	*
United Kingdom	5	234,522	2.1	(0.3 - 4.0)	0.1	*	*	*	*	*

Source: Los Angeles County Department of Public Health, TB Control Program & Public Use Microdata Survey via IPUMS. Isoniazid resistance is not exclusive of other resistance.

† Incidence limits calculated based on standard error of denominator which in turn was calculated based on PUMS replicate weights.

**For diagnosis year 2005, excluded cases indicated to be homeless, incarcerated or in long-term care facilities because ACS 2005 excluded these populations from group quarters.*

Table 3: TB Incidence Rates among Foreign-born Persons by Demographic Characteristic, Los Angeles County 2005*-2011

Demographic Characteristic		Diagnoses	Person-Years	Unadjusted Incidence per 100,000	95% Confidence Interval	Percentage of Total Diagnoses
Age						
	0 – 19	118	1,787,626	6.6	(5.4 - 7.8)	3%
	20 – 39	1,098	9,085,859	12.1	(11.4 - 12.8)	28%
	40 – 59	1,365	9,597,152	14.2	(13.5 - 15.0)	35%
	60 – 79	976	3,782,404	25.8	(24.2 - 27.4)	25%
	80 – 106	389	784,359	49.6	(44.7 - 54.5)	10%
Gender						
	Male	2,296	12,252,800	18.7	(18.0 - 19.5)	58%
	Female	1,650	12,784,600	12.9	(12.3 - 13.5)	42%
Years in Residence						
	0 – 1	576	824,925	69.8	(64.1 - 75.5)	15%
	2 – 4	462	1,774,188	26.0	(23.7 - 28.4)	12%
	5 – 9	579	3,316,024	17.5	(16.0 - 18.9)	15%
	10 – 19	881	6,823,639	12.9	(12.1 - 13.8)	22%
	20 – 93	1,448	12,298,624	11.8	(11.2 - 12.4)	37%
Year of Diagnosis						
	2005	652	3,593,316	18.1	(16.8 - 19.5)	17%
	2006	636	3,569,735	17.8	(16.4 - 19.2)	16%
	2007	597	3,642,877	16.4	(15.1 - 17.7)	15%
	2008	554	3,538,054	15.7	(14.4 - 17.0)	14%
	2009	501	3,567,900	14.0	(12.8 - 15.3)	13%
	2010	507	3,553,789	14.3	(13.0 - 15.5)	13%
	2011	499	3,571,729	14.0	(12.7 - 15.2)	13%

**For diagnosis year 2005, excluded cases indicated to be homeless, incarcerated or in long-term care facilities because ACS 2005 excluded these populations from group quarters.*

Table 4: TB Incidence Rates and Incidence Rate Ratios by Selected Country of Birth, Los Angeles County 2005*-2011

Country of Birth	Unadjusted			Adjusted Poisson GLM†			Adjusted Negative Binomial GLM†		
	Incidence	IRR	Standard	IRR	Standard	Robust	IRR	Standard	Robust
Philippines	44.3	4.61	(4.28 - 4.94)	4.35	(3.91 - 4.83)	(3.83 - 4.94)	4.23	(3.71 - 4.82)	(3.70 - 4.83)
Vietnam	38.7	4.02	(3.54 - 4.51)	4.54	(3.93 - 5.25)	(3.84 - 5.36)	4.49	(3.79 - 5.31)	(3.76 - 5.34)
India	29.6	3.08	(2.47 - 3.69)	2.48	(2.00 - 3.09)	(1.93 - 3.19)	2.45	(1.94 - 3.09)	(1.90 - 3.14)
China	27.7	2.88	(2.53 - 3.23)	2.07	(1.79 - 2.39)	(1.74 - 2.46)	2.00	(1.69 - 2.35)	(1.68 - 2.37)
Korea	22.2	2.31	(2.02 - 2.59)	2.28	(1.97 - 2.64)	(1.94 - 2.68)	2.20	(1.86 - 2.60)	(1.86 - 2.60)
Guatemala	17.6	1.83	(1.58 - 2.07)	2.16	(1.85 - 2.52)	(1.78 - 2.61)	2.08	(1.74 - 2.48)	(1.72 - 2.51)
Mexico	12.4	1.29	(1.22 - 1.37)	1.70	(1.55 - 1.87)	(1.52 - 1.90)	1.67	(1.47 - 1.89)	(1.48 - 1.87)
Other foreign country	9.6	ref		ref			ref		
El Salvador	9.4	0.98	(0.83 - 1.13)	1.26	(1.06 - 1.49)	(1.05 - 1.51)	1.23	(1.01 - 1.48)	(1.02 - 1.47)
Dispersion statistic	N/A			1.54			1.3		
AIC	N/A			6907			6819		

Source: Los Angeles County Department of Public Health, TB Control Program & Public Use Microdata Survey via IPUMS.

**For diagnosis year 2005, excluded cases indicated to be homeless, incarcerated or in long-term care facilities because ACS 2005 excluded these populations from group quarters.*

†*Adjusted model includes the following covariates: age, gender, length of residence and year of diagnosis.*

Table 5: TB Incidence Rate Ratios by Selected Demographic Characteristics – Bayesian Models, Los Angeles County 2005*-2011

Demographic Characteristics		Poisson Bayes	Negative Binomial Bayes	Negative Binomial Bayes with informative priors on population estimates
Country of Birth				
	Philippines	4.35 (3.88 - 4.84)	4.22 (3.67 - 4.82)	4.26 (3.70 - 4.86)
	Vietnam	4.54 (3.88 - 5.25)	4.47 (3.70 - 5.32)	4.41 (3.68 - 5.26)
	India	2.48 (1.96 - 3.08)	2.43 (1.88 - 3.08)	2.42 (1.86 - 3.07)
	China	2.07 (1.76 - 2.39)	1.99 (1.66 - 2.36)	2.01 (1.68 - 2.37)
	Korea	2.28 (1.94 - 2.64)	2.19 (1.83 - 2.60)	2.22 (1.85 - 2.63)
	Guatemala	2.15 (1.82 - 2.52)	2.07 (1.70 - 2.49)	2.05 (1.69 - 2.46)
	Mexico	1.71 (1.54 - 1.88)	1.67 (1.45 - 1.90)	1.66 (1.46 - 1.88)
	El Salvador	1.26 (1.05 - 1.49)	1.22 (0.99 - 1.49)	1.22 (0.99 - 1.48)
	Other foreign country	reference	reference	reference
Age				
	0 – 19	0.18 (0.15 - 0.22)	0.19 (0.15 - 0.23)	0.19 (0.15 - 0.23)
	20 – 39	0.56 (0.51 - 0.61)	0.61 (0.54 - 0.68)	0.60 (0.54 - 0.67)
	40 – 59	reference	reference	reference
	60 – 79	1.91 (1.74 - 2.08)	1.95 (1.74 - 2.18)	1.94 (1.73 - 2.16)
	80 – 106	4.28 (3.77 - 4.81)	4.39 (3.76 - 5.07)	4.32 (3.72 - 4.96)
Gender				
	Male	reference	reference	reference
	Female	0.62 (0.58 - 0.66)	0.64 (0.59 - 0.69)	0.64 (0.59 - 0.69)
Years in Residence				
	0 – 1	10.54 (9.41 - 11.73)	10.93 (9.53 - 12.46)	10.65 (9.26 - 12.17)
	2 – 4	3.88 (3.45 - 4.35)	3.85 (3.33 - 4.42)	3.89 (3.36 - 4.47)
	5 – 9	2.51 (2.25 - 2.79)	2.51 (2.19 - 2.86)	2.51 (2.20 - 2.85)
	10 – 19	1.56 (1.42 - 1.71)	1.62 (1.43 - 1.81)	1.61 (1.43 - 1.80)
	20 – 93	reference	reference	reference
Year of Diagnosis				
	2005	reference	reference	reference
	2006	1.01 (0.90 - 1.13)	1.01 (0.87 - 1.17)	1.00 (0.86 - 1.15)
	2007	0.90 (0.80 - 1.01)	0.93 (0.80 - 1.07)	0.91 (0.78 - 1.06)

	2008	0.86 (0.76 - 0.96)	0.84 (0.71 - 0.97)	0.84 (0.71 - 0.97)
	2009	0.80 (0.70 - 0.90)	0.78 (0.66 - 0.91)	0.78 (0.67 - 0.91)
	2010	0.79 (0.69 - 0.89)	0.75 (0.64 - 0.87)	0.76 (0.64 - 0.88)
	2011	0.79 (0.69 - 0.89)	0.76 (0.64 - 0.88)	0.76 (0.65 - 0.89)

Source: Los Angeles County Department of Public Health, TB Control Program & Public Use Microdata Survey via IPUMS.

**For diagnosis year 2005, excluded cases indicated to be homeless, incarcerated or in long-term care facilities because ACS 2005 excluded these populations from group quarters.*

†*Adjusted model includes the following covariates: age, gender, length of residence and year of diagnosis.*

Spatial Distribution of TB Incidence, Los Angeles County 2005-2001

Abstract

Introduction

In Los Angeles County, the tuberculosis (TB) disease incidence rate is seven times higher among foreign-born persons than U.S.-born persons and varies substantially by country of birth [75]. Spatial analyses can be used to identify areas of high TB disease incidence which may be helpful in scaling up latent TB testing.

Methods

Data on 5,447 diagnosed TB cases from the Los Angeles County Department of Public Health TB Control Program were combined with stratified population estimates available from the Public Use Microdata Survey to calculate TB incidence rates for the years 2005 through 2011. Country-specific TB disease incidence rates were calculated and naïve smoothing was applied. We investigated residual spatial component when modelling with country of birth only, all covariates and all covariates plus non-spatial error.

Results

There were notable differences in the unadjusted and spatially-smoothed maps of TB disease incidence for selected high-incidence countries, namely Mexico, Vietnam and the Philippines. Spatially-smoothed maps showed areas of high incidence in downtown Los Angeles and surrounding areas for both Philippines-born and Vietnam-born persons. Areas of high

incidence were more dispersed for Mexican-born persons. Residual spatial features in models incorporating all covariates and non-spatial error suggested that the spatial distribution of the disease cannot be full explained using the available covariates.

Conclusion

This study highlights areas of high TB incidence within Los Angeles County both for U.S.-born cases and for cases born in Mexico, Vietnam and the Philippines. It also highlights areas that were high incidence even when accounting for non-spatial error and important covariates including age, sex, and years in residence. The spatial patterning provided in the maps provide complementary granularity to descriptions of the local disease burden which may help inform the continued public health response by supporting targeted testing and focusing local efforts to support new recommendations from the USPSTF regarding testing for latent TB infection in high-risk individuals.

Introduction

Los Angeles County (LAC) is a capacious jurisdiction covering 4,058 square miles with a large, diverse population of more than 10 million people, of which 3.5 million are foreign-born [27]. TB disease incidence is notably higher among foreign-born persons and the incidences rate by country of birth have been described [39, 75, 76]. However, the spatial distribution of TB incidence in Los Angeles County has largely remained unreported. Local health departments do not routinely report TB disease spatially although case address is reportable by law; some departments have done spatial analyses as special projects [77-79].

We anticipate that TB disease is clustered in hotspots given spatial analyses conducted in other locales and the nature of TB transmission and reactivation. TB transmission and reactivation often reflect social patterning which is, in part, spatially defined. Furthermore, it is assumed that, even when accounting for known and available risk factors, TB disease is unlikely to be evenly distributed. Information on areas of elevated TB disease are especially relevant now as there has been substantial effort at local, state and national levels to scale up latent TB infection testing as part of targeted testing and treatment. The United States Preventive Services Task Force (USPSTF) recently issued a grade B recommendation for latent TB infection testing in high-risk individuals [80]. An important consequence of this recommendation is that, under current ACA regulations, health insurance plans would be required to cover the cost of latent TB infection testing.

We used data from the LAC TB surveillance system and the American community survey to produce unadjusted TB incidence by country of birth and sub-county area. We then smoothed these maps using empirical Bayes to attenuate the effect of sparse data. Finally, we created extended models to account for additional covariates and non-spatial error.

Methods

Data collection and management have been described previously [75]. Briefly, between 2005 and 2011, 5,447 TB cases meeting the definition for the report of a verified case of tuberculosis (RVCT) were diagnosed and reported to the Los Angeles County Department of Public Health TB Control Program TB surveillance system [14, 81]. Address at diagnosis was geocoded using a Los Angeles County-approved geolocator allowing the case to be assigned to one of 67 Public Use Microdata Areas (PUMAs) in Los Angeles County as defined by the 2000 U.S. Census [42, 82]. Data for population estimates stratified by PUMA and other covariates of interest were obtained from the Integrated Public Use Microdata Series, a curated copy of the U.S. Census's Public Use Microdata Survey and other microdata [26]. PUMAs were chosen as the geography of interest because they were the smallest area for which the full joint distribution for key covariates was available. Population estimates were calculated using replicate weights; exclusions, replicate weights and sampling frame were discussed in detail in prior work [45, 75]. In summary, 494 (9%) cases were excluded due to missing data or differences in sampling frame between Los Angeles County TB surveillance and the American Community Survey leaving 4,953 cases available for analysis. An additional 1,008 (18.5%) U.S.-born cases were excluded from those analyses which included years in residence as a covariate because years in residence was undefined for U.S.-born cases and regardless collinear with age, an important covariate [75].

Unadjusted TB incidence rates stratified by PUMA alone and by country of birth and PUMA were calculated. Adjustment was achieved using multiple regression in a Bayesian framework and a conditional autoregressive regressive term from Besag et al. which is used frequently in spatial applications [83-85]. The preliminary model (Equation 1 below) accounts for area and country of birth only. Following the notation of Kleinschmidt et al., Y_{ic} is defined as the observed diagnoses occurring in area i and among country of birth c; P_{ic} is defined as the person-time for the same stratum [86]. Additionally, we define $\eta_{ic} \equiv E[Y_{ic}]$ and assume that $Y_{ic} \sim Poisson(\eta_{ic})$. The transformed linear regression is then:

$$log(\eta_{ic}) = \log(P_{ic}) + \alpha + \beta_c X_c + \varphi_i \#(1)$$

where φ_i denotes a spatially-correlated random effects term defined by the following [85]:

$$\varphi_i | \varphi_{-i} = N\left(\bar{\varphi_\iota}, \frac{{\sigma_\varphi}^2}{n_i}\right)$$

$$\bar{\varphi_\iota} = \frac{1}{n_i} \sum_{j \in neighbors\ of\ i} \varphi_i$$

Neighbors of area i were defined with queen-style contiguity (Figure 1) [87].

Subsequent models (equations 2 and 3) used the following transformed linear regressions:

$$log(\eta_{is}) = \log(P_{is}) + \alpha + \boldsymbol{\beta X} + \varphi_i \#(2)$$

$$log(\eta_{is}) = \log(P_{is}) + \alpha + \boldsymbol{\beta X} + \varphi_i + \omega_s \#(3)$$

where s denotes the stratum, $\boldsymbol{\beta X}$ denotes the vectors of covariates and covariate betas, ω_s denotes the spatially-uncorrelated heterogeneity with the distribution $\omega_s \sim N(0, \sigma^2)$. The priors were set as follows: α was given a flat prior, $\boldsymbol{\beta}$ were given N(0, 1000), and φ_i and ω_s were both given Gamma(0.5, 2000). Bayesian models were run with two chains for 100,000 iterations and 10,000 iterations of burn-in. Mixing was evaluated through visual inspection of caterpillar plots and density charts. ArcGIS 10.0 was used to geocode and check geolocation. R version 3.4, R Studio version 1.0.143 and variety of packages were used to manage and analyze data and create maps [51, 55-62, 64, 65, 88-93]. Bayesian models were run in OpenBUGS version 3.2.2 rev 1012 [52] (see Appendix 3 for OpenBUGS model code). Due to limitations stemming from sparse data for most country-of-birth groups, only a select group of countries of birth were analyzed via unadjusted and adjusted TB incidence (Equation 1). Data from all country-of-birth groups were included in subsequent models (Equations 2 and 3).

Results

As previously reported, the tuberculosis incidence rate in Los Angeles County 2005-2011 was 7.2 per 100,000, with 2.3 per 100,00 occurring among U.S.-born persons and 15.8 per 100,000 occurring among foreign-born persons [75]. The map for unadjusted incidence among all residents shows higher incidence in central areas of the county and lower incidence in outer areas (Figure 2A). For reference, the California and U.S. TB disease incidence rate in the same period were 7.1 per 100,00 and 4.1 per 100,000 [94]. Areas of notable high incidence include Panorama City, Pico Heights and Echo Park, and Monterey Park-Rosemead, which are in the Northwest, center and East sections of the county. These areas had unadjusted incidences of 13.2, 19.7, 17.2 and 19.2 per 100,000 respectively. Adjustment through Bayesian smoothing, using Equation 1, had minimal effect on estimates (Figure 2B); median absolute difference between adjusted and unadjusted incidences was 0.13 per 100,000 with a maximum of 0.59 per 100,000.

TB incidence among U.S.-born persons and foreign-born persons showed different spatial patterns. Among U.S.-born persons, there were areas of high incidence including Los Angeles City downtown, Watts and East Los Angeles (Figure 2C). These areas had incidences of 8.0, 7.9 and 6.1 per 100,000 respectively. In contrast, among foreign-born persons, TB incidence was notably higher in Monterey Park/Rosemead, Pico Heights and Echo Park. These areas had TB incidences of 32.8, 26.2 and 28.3 per 100,000 respectively. Also noteworthy were two areas of elevated incidence separated from the central form, specifically Panorama City (northwest) that had an incidence rate of 23.1 per 100,000 and Carson (due South of downtown) that had an incidence rate of 25.1 per 100,00. Changes in estimates via Bayesian smoothing for both U.S.-born and foreign-born persons were minor (Figure 2D, Figure 2F).

Prior reports have shown notable differences in incidence by country of birth with the largest absolute number of cases occurring among persons born in Mexico, Philippines and Vietnam [39, 75]. The map for unadjusted incidence rates among Mexican-born persons shows a condensed spatial form centered north of downtown Los Angeles in contrast to maps for unadjusted incidence among Filipino-born and Vietnamese-born persons which show more dispersed patterning throughout the county (Figure 3: A, C, E). Maps of incidence rates adjusted through smoothing had a less dispersed pattern than unadjusted maps and show concentrated areas of high incidence (Figure 3: B, D, and F). Maps for adjusted incidence rates among Mexican-born and Filipino-born persons show a cluster of areas of high incidence centered on the Los Angeles City downtown (Figure 3: B, D). The adjusted map for incidence rates among Vietnamese-born persons shows a small area of high incidence centered on the Los Angeles downtown (Figure 3: F).

Subsequent models including additional covariates (Equations 2 and 3) showed condensed spatial patterns (Figure 4). Models are country of birth only, with all covariates and with all covariates and a non-spatial error term. Maps show a high degree of spatial patterning even when accounting for covariates and when accounting for covariates and non-spatially correlated error.

Discussion

TB disease incidence has a distinctive spatial pattern overall and in several country-of-birth sub-groups, with several identifiable hotspots within Los Angeles County. Areas of elevated incidence among U.S.-born persons were evident in downtown Los Angeles as well as to the East of the city center. Among Filipino-born and Vietnamese-born persons, unadjusted TB incidence exhibits a highly-dispersed spatial pattern. In contrast, among Mexican-born persons, a condensed spatial pattern in unadjusted TB incidence is evident. Maps of TB incidence rates adjusted through smoothing show, at least in the case of incidence among Filipino-born persons, a strong spatial pattern with areas of high incidence centered on the Los Angeles downtown area. The adjusted map of incidence among Vietnamese-born persons shows one area of high disease incidence. The notable differences in the unadjusted and adjusted maps for the selected countries of birth shows the utility of empirical Bayesian smoothing; low absolute values of strata numerators and denominators for Filipino-born and Vietnamese-born subgroups produced highly variable incidence estimates. Smoothed maps are easier to interpret because they account for part of the underlying uncertainty in the incidence rates.

Spatial patterning persists even when considering covariates and non-spatial error. Spatial patterns of the single covariate models (country of birth) suggests that we are justified in constructing further models to explain the spatial differences, in that country of birth alone is not sufficient to explain the existing spatial pattern. This is confirmed by examining the spatial distribution of the incidence rate ratio for the country of birth only model (Figure 4A). However, additional models attenuate but do not remove the spatial heterogeneity and form as is evidenced by Figure 4B and Figure 4C, suggesting that additional data is necessary to explain the clustering of high incidence areas. We believe this confirms the complex mixture of recent transmission and reactivation of latent infection that is ongoing within Los Angeles County and the nation as a whole. Both recent transmission and reactivation has socio-spatial components so it is difficult to disentangle which of these is underlying the clusters seen in adjusted results. Additional analysis incorporating recent advancements in prediction of recent transmission would helpful in understanding which of the case sub-populations is driving clustering[71].

For the local clinical community, we believe that this information can add supportive detail to a clinical risk assessment. The California Department of Public Health TB Control Branch recently issued a tuberculosis risk assessment [70]. Additional detail on country of birth specific risks and even risks specific to a local community could help providers though care must be taken that this information does not divert attention from or discourage testing among other high-risk groups.

This analysis has additional limitations beyond issues of cases ascertainment and survey error discussed in prior work [75]. This analysis is vulnerable to the modifiable areal unit problem (MAUP) and may yield different results based on the size and shape of the areas under study. Low absolute numbers in strata numerators and denominators make incidence calculations more variable. Edge effects mean that PUMAs on the edge of the county have fewer neighbors and so may not be as well smoothed as PUMAs in the middle of the county. Also, cases from Long Beach and Pasadena are not included here. As a result, areas around Pasadena are missing a neighbor and so are not smoothed as they would be if Pasadena cases had been included. Similarly, areas around Long Beach are also not smoothed as they would be; this is especially problematic as Long Beach had a substantial number of cases for the size of its population. PUMA boundary definitions from the 2000 Census allowed for non-contiguous areas. In Los Angeles County, there are several non-contiguous PUMAs which can distort smoothing process by creating neighbors for non-contiguous areas.

Conclusion

This study confirms that TB disease incidence is spatially heterogeneous within Los Angeles County and remains so within high-risk countries of birth and when accounting for non-spatial error and important covariates including age, sex, and years in residence. The spatial patterning in the maps provides complementary information to descriptions of the local disease burden. We hope that this information will inform the continued public health response by supporting targeted testing and focusing local efforts to support new recommendations from the USPSTF regarding testing for latent TB infection in high-risk individuals. Because TB control has historically relied on direct intervention on transmission through case investigation and treatment, therefore analyses like this one may have application to TB control efforts, either through influencing case investigation or alternative control efforts. We expect that these analyses could be extended by using ecological variables and by integrating information from algorithms to identify recent transmission [71].

This study reinforces the importance of spatial data in local description and suggests further that they be of use in predictive models of TB incidence, both directly, as a covariate, and indirectly, through leveraging of other spatial data. For example, domestic TB prediction models based on routinely-reported TB surveillance data lack socio-economic status and crowding data. Spatial data could augment existing TB data by linking available microdata or ecological data to reported cases.

Figure 1: Queen Neighbor Matrix for PUMAs from Census 2000, Los Angeles County with Long Beach and Pasadena Removed

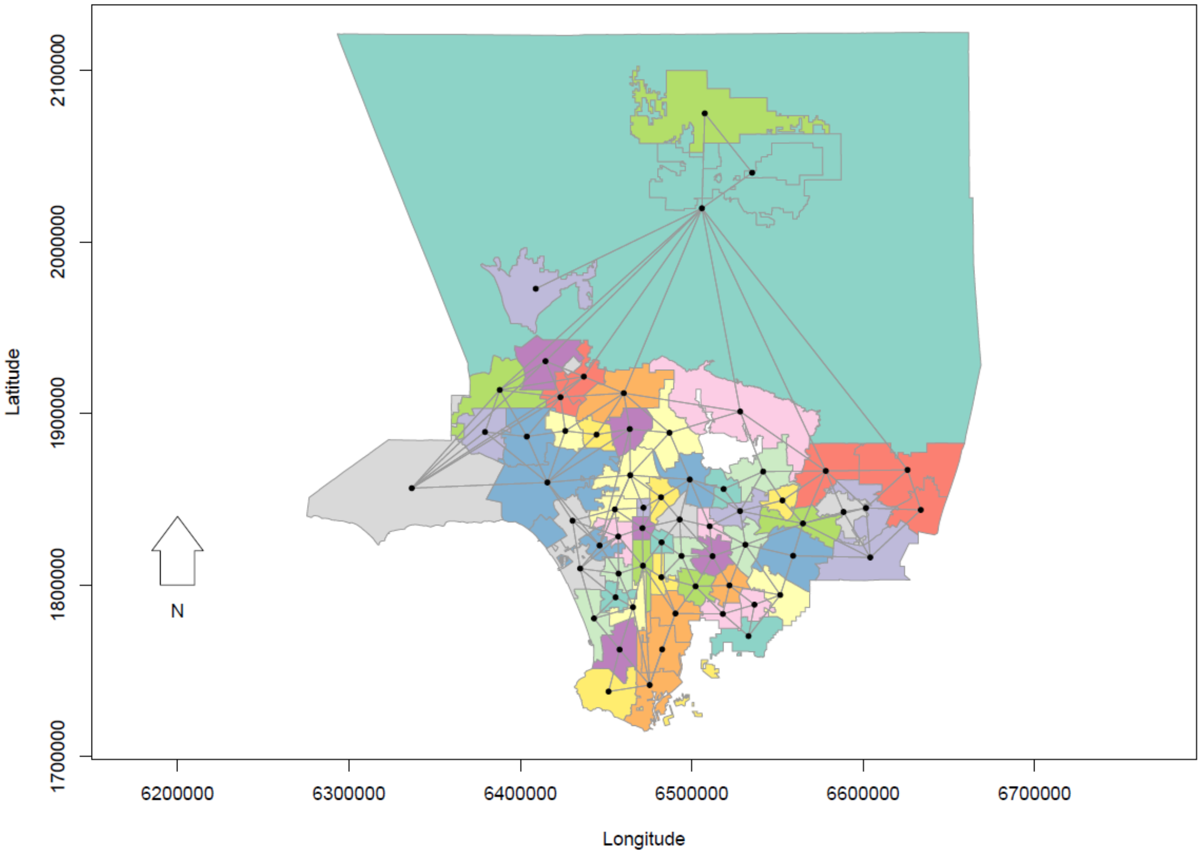
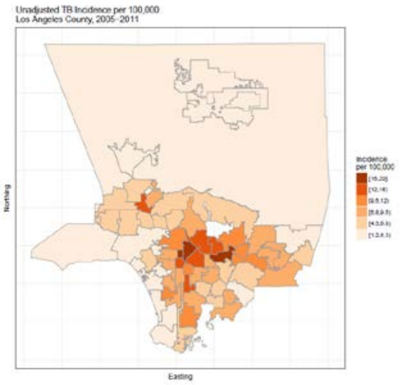
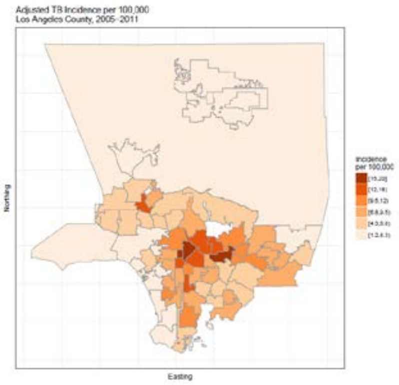
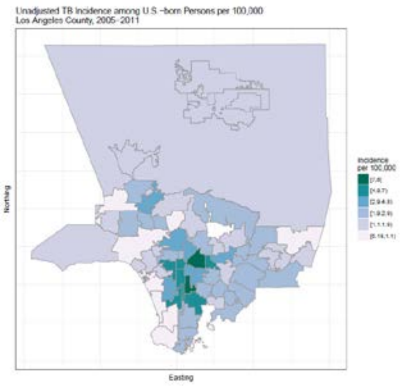
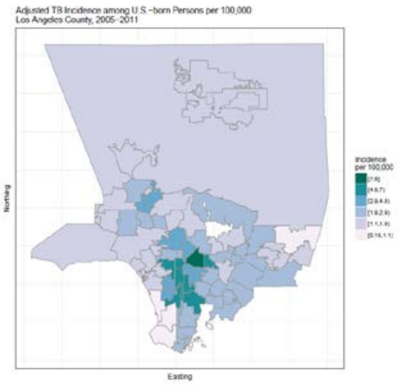


Figure 2: Unadjusted and Adjusted TB Incidence among Selected Subgroups, Los Angeles County 2005-2011.

	Unadjusted	Adjusted
All Residents	 Unadjusted TB Incidence per 100,000 Los Angeles County, 2005–2011 Northing Easting Incidence per 100,000 [15,22] [12,15) [9.5,12) [6.8,9.5) [4.3,6.8) [1.2,4.3) A	 Adjusted TB Incidence per 100,000 Los Angeles County, 2005–2011 Northing Easting Incidence per 100,000 [15,22] [12,15) [9.5,12) [6.8,9.5) [4.3,6.8) [1.2,4.3) B
US born	 Unadjusted TB Incidence among U.S.–born Persons per 100,000 Los Angeles County, 2005–2011 Northing Easting Incidence per 100,000 [7,9] [4.8,7) [2.9,4.8) [1.9,2.9) [1.1,1.9) [0.16,1.1) C	 Adjusted TB Incidence among U.S.–born Persons per 100,000 Los Angeles County, 2005–2011 Northing Easting Incidence per 100,000 [7,9] [4.8,7) [2.9,4.8) [1.9,2.9) [1.1,1.9) [0.16,1.1) D

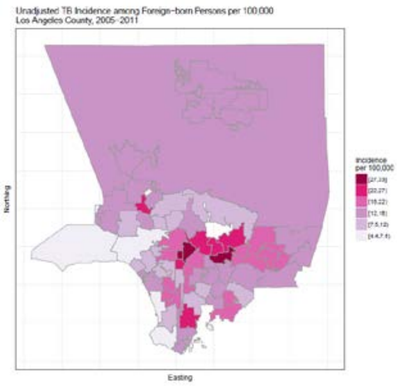
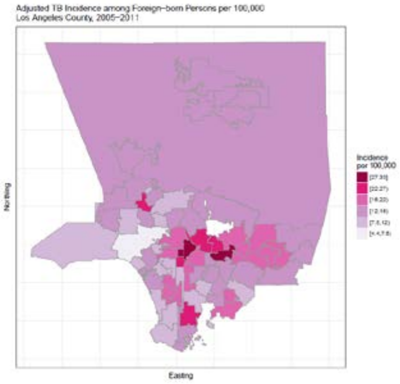
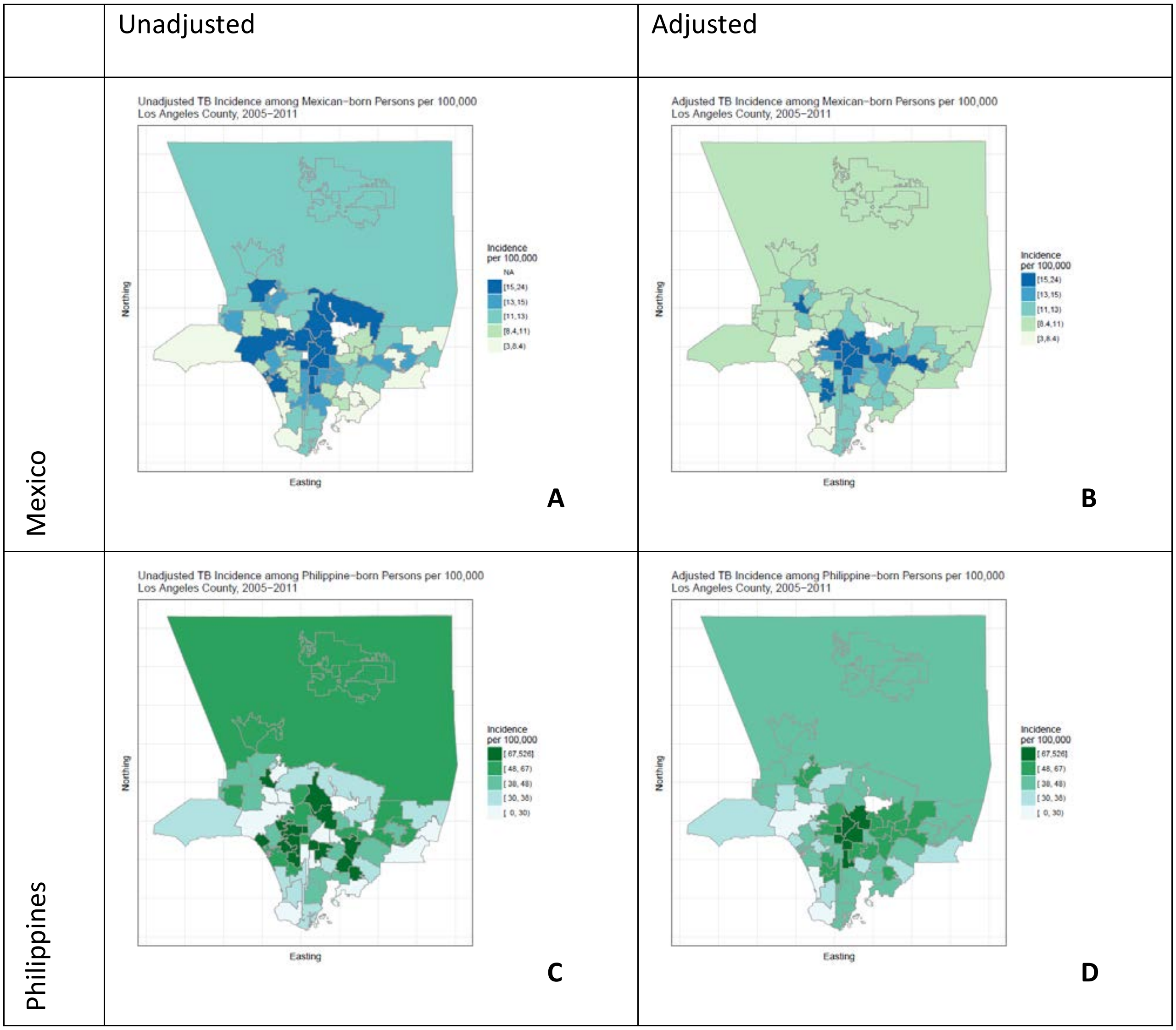
Foreign born	 Unadjusted TB Incidence among Foreign−born Persons per 100,000 Los Angeles County, 2005−2011 Northing Easting Incidence per 100,000 [27,33] [22,27) [18,22) [12,18) [7.5,12) [4.4,7.5) **E**	 Adjusted TB Incidence among Foreign−born Persons per 100,000 Los Angeles County, 2005−2011 Northing Easting Incidence per 100,000 [27,33] [22,27) [18,22) [12,18) [7.5,12) [4.4,7.5) **F**

Figure 3: Unadjusted and Adjusted TB Incidence among Selected Countries of Birth, Los Angeles County 2005-2011.



Unadjusted TB Incidence among Mexican-born Persons per 100,000 Los Angeles County, 2005–2011

Adjusted TB Incidence among Mexican-born Persons per 100,000 Los Angeles County, 2005–2011

Unadjusted TB Incidence among Philippine-born Persons per 100,000 Los Angeles County, 2005–2011

Adjusted TB Incidence among Philippine-born Persons per 100,000 Los Angeles County, 2005–2011

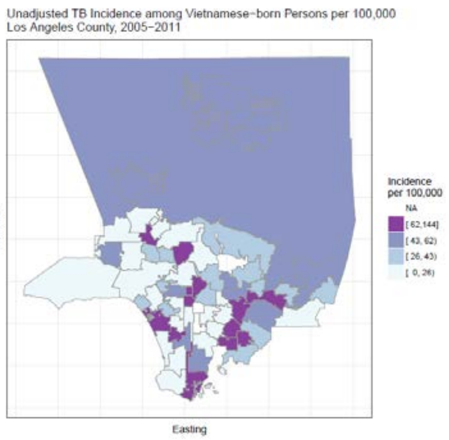
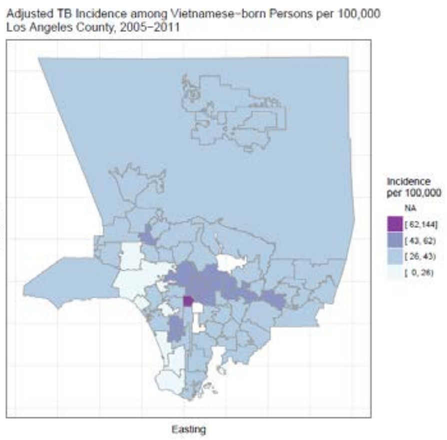
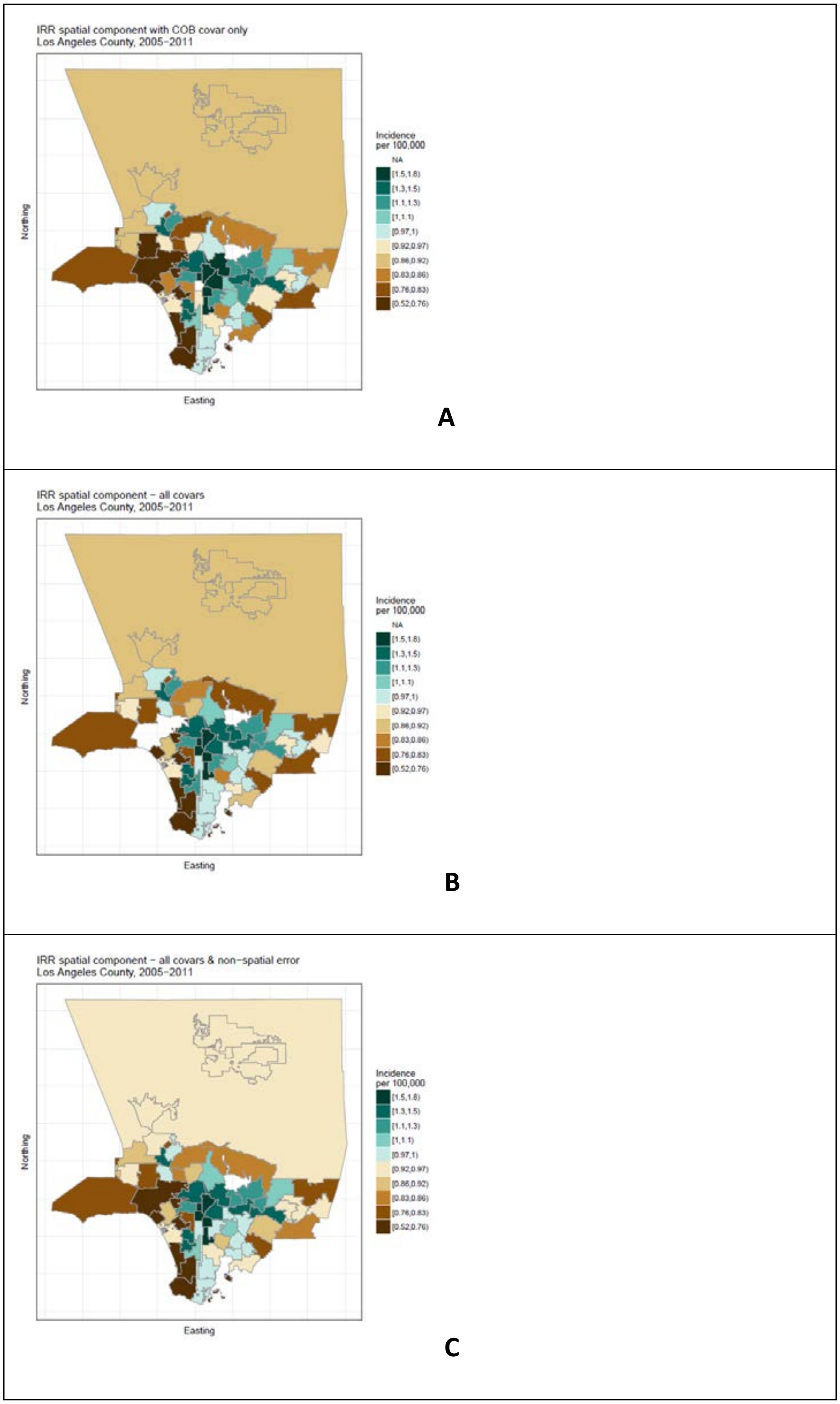
Vietnam	 Unadjusted TB Incidence among Vietnamese-born Persons per 100,000 Los Angeles County, 2005-2011 Northing Easting Incidence per 100,000 NA [62,144] [43, 62) [26, 43) [0, 26) E	 Adjusted TB Incidence among Vietnamese-born Persons per 100,000 Los Angeles County, 2005-2011 Northing Easting Incidence per 100,000 NA [62,144] [43, 62) [26, 43) [0, 26) F

Figure 4: Incidence Rate Ratio of Spatial Component of Full Bayesian models



TB Disease Incidence Rate Prediction among Foreign-born Persons, Los Angeles County 2005-2011

Abstract

Introduction

The TB disease incidence rate among foreign-born persons in Los Angeles County is seven times the rate among US-born [75]. Detailed description of this epidemic has helped focus local control efforts, but further refinement in the prediction of TB diagnoses would provide additional benefit. Earlier efforts have shown that area-based ecological proxies of known risk factors for TB diagnosis improve predictive models. Here we attempt to improve prediction of previously-described non-spatial and spatial models.

Methods

Strata-specific incidence rates were calculated using TB case data from the LA County DPH TB surveillance system combined with stratified population estimates constructed using the Public Use Microdata Series (PUMS). Using a framework of known risk factors, we selected comparable area-based, ecological variables estimated from PUMS. AIC-monitored backward elimination of ecologic covariates in a non-spatial Poisson general linear model was used with data from a training partition. Using a testing partition, the ecologic model was compared with the non-ecological model with mean squared error (MSE). Using a Bayesian spatial model and a 10-fold partitioning scheme, we ran stochastic variable selection to test the inclusion of ecologic covariates and tested the most supported spatial model against reversed test partition.

Results

No improvement was shown by the inclusion of ecologic covariates selected by the bootstrap backward elimination. The MSE of the minimum model was 0.214 and the MSE of the model including ecologic covariates was 0.208. Using stochastic variable selection in a spatial context, the model most select excluded all ecological covariates. The improvement of MSE across all test partitions was minimal.

Conclusion

Area-specific ecological variables did not improve prediction in non-spatial or spatial models. Additional work is necessary to test whether prediction is improved with ecological variables attach to other covariates namely country of birth. Models using country of birth, age, sex and years in residence have reasonable predictive ability for TB diagnoses among the foreign-born. Changes to routine reporting may be necessary to examine socio-economic status among TB diagnoses.

Introduction

The TB incidence rate among foreign-born persons is Los Angeles County is seven times higher than TB incidence among U.S.-born [75]. Among some country of birth groups, the incidence rate is more than 22 times the rate among U.S.-born [75]. Increased incidence rates among country of birth subgroups have been noted both locally and nationally [39, 95]. Earlier efforts to describe TB epidemics in detail have been fruitful, but were designed to provide accurate estimation not prediction.

Analyses of communicable diseases are hampered by violations of the distributions on which standard regression models rely. This has led many researchers to use compartment models or agent-based models for prediction and estimation of communicable disease [96, 97]. However, the TB epidemic has become perhaps easier to predict because recent analyses in the U.S. suggest 85% of cases are due to reactivation with the remaining of latent TB infection and with the remaining 15% due to transmission [71].

Much is known about the individual-level and ecological-level risk factors for TB disease, yet it is unclear whether earlier estimation models recast as prediction models can be improved especially in the domestic context [75, 98-101]. Furthermore, there is a need to consider the utility to public health practitioners at the state and local level [102, 103].

Here we attempt to assess the predictive ability of previously-described non-spatial and spatial estimation models and improve these models with ecological data [98, 101]. Using a framework on known risk factors, we selected reasonable analogues to these risk factors from comparable area-based, ecological variables. For non-spatial models, we reduced the available covariates using a backward elimination. For spatial models, we conducted stochastic variable selection (SVS) in a Bayesian spatial model.

Methods

Data Sources

Data on TB diagnoses drawn from the Los Angeles County Department of Public Health TB surveillance system were combined with stratified population estimates constructed using the Public Use Microdata Series (PUMS) available through the Minnesota Population Center's Integrated Public Use Microdata Series (IPUMS) as described previously [26, 42, 75, 101]. Ecological data was also drawn from PUMS via IPUMS. Addresses at diagnoses was geocoded and assigned to an U.S. Census 2000 Public Use Microdata Area (PUMA) using a written protocol [75].

Construction of Prediction Model: Two Approaches

Construction of prediction models is categorically different from the construction of estimation or causal models, both in aims and challenges. Causal models are designed to elucidate causal pathways for a specific outcome and estimate parameters describing those pathways. In contrast, prediction models are designed to optimize the prediction of an event without regard to establishing causal links. The canonical issues in constructing prediction models are over-fitting and limitations of computational resources [104]. Over-fitting occurs when the model is constructed in such a way as to be highly-predictive with the existing data, but minimally-predictive with new data. A slew of methods have been developed to overcome these hurdles; here, we use two [104, 105]. To construct the non-spatial model, we created bootstrap replicates and used a stepwise elimination method on a negative binomial general linear model (GLM) for each replicate, similar to prior work [75]. For the spatial model, we used stochastic variable selection (SVS) on a Bayesian spatial Poisson general linear mixed model (GLMM) described previously [101, 106]. [1,19]

Ecological Variables

To reduce computational burden and improve performance of variable selection procedures, we selected only those PUMS variables that were reasonable analogues of TB risk factors detailed in a framework by Hargreaves et al. [100]. A total of 14 variables were selected and, using survey weights, 15 population estimates were calculated for each PUMA (two versions of income were created): proportion renting home, proportion enrolled in the Supplemental Nutrition Assistance Program (SNAP) commonly called food stamps, proportion linguistically isolated, average number of rooms, proportion with incomplete plumbing, average number of units in structure, average family size, proportion with a high school education (or equivalent) or higher, median income, average income, average Federal poverty level, average socio-economic index, average Hauser and Warren socio-economic index, proportion with any health insurance, population density. Households in which “no person age 14+ speaks only English at home, or no person age 14+ who speaks a language other than English at home speaks English ‘Very well’…” are categorized as linguistically isolated; variable definitions are available from

IPUMS and other sources [26]. Population density was calculated by estimating the population by PUMA using survey weights for the entire study period and dividing by the number of years in the study period and the number of square miles in the associated PUMA. Our criteria for what constituted a reasonable analogue were subjective, but given our goal of constructing a prediction model, we were generally inclusive. Later analysis assessed the predictive value of each variable and managed issues arising from variable correlation. Certain variable pairs were known *a priori* to be correlated, for example median income and proportion with high school education or higher as well the socio-economic index and the Hauser and Warren socio-economic index, but both members of the correlated pairs were included because statistical methods could better determine which member of the pair (if either) was more predictive of the outcome. Note that some variables were assessed at the individual level and others at the household level. Household level attributes, such as average number of units in structure, were attached to individuals. All ecological variables were estimated for the study period except for proportion with any health insurance which had data for 2008-2011 only. A total of 5,447 cases were drawn from the surveillance system with diagnosis dates between 2005 and 2011. We excluded 1,008 (19%) U.S.-born cases and 494 (9%) cases due to missing data or definition incompatible with denominator data: 26 cases with missing or unknown country of birth, 20 cases that were in correctional facilities (or missing this variable) and diagnosed in 2005, 16 cases that were in long-term care (or missing this variable), 277 cases that were homeless (or missing this variable) and not housed in group quarters, 13 cases diagnosed in 2005 that were homeless and living in group quarters, 14 administrative cases, 112 cases that were foreign-born but missing years in residence, 15 cases with countries of birth not included in the PUMS data and 1 case that could not be assigned to a PUMA. Additional cases were excluded in spatial analysis, leaving 3,792 cases available. Cases residing in Long Beach or Pasadena were not included as those cases were reported to the Pasadena City Health Department and the Long Beach Health and Human Services, respectively.

Non-spatial Prediction with Bootstrap Replicates

After selecting analogues to TB risk factors from PUMS data, two methodological approaches were pursued. The first approach constructed a non-spatial prediction model. Cross-correlation of ecological variables was examined via correlation plots and a variance cluster tree [107, 108]. Redundancy among ecological variables was assessed using parametric additive models available through the redun function of the Hmisc package [107]. Variables were eliminated stepwise. In each round, R^2 was calculated for each variable based all other variables and the variable with the highest R^2 was eliminated. The process was terminated when all variable predictions were below a pre-determined cutoff of $R^2 = 0.9$.

Individual-level diagnosis data were partitioned into a training dataset with two-thirds of the data and testing dataset with the remaining third. Likewise, denominators were calculated for training and testing datasets by multiplying the stratum-specific denominator by 2/3 and 1/3 respectively. Ecological variables selected with the redundancy test were entered into a negative binomial general linear model (GLM) together with covariates shown to be predictive in previous work, namely age, sex, country of birth and years in residence [39, 40, 75]. The base model was defined as containing the covariates age, sex, country of birth and years in residence. This model was then subjected to a stepwise backwards elimination process using the Akaike Information Criteria (AIC) on 1000 bootstrap samples of the training dataset. This was achieved using the a custom bootstrap function and the stepAIC functions from the MASS packages [109]. To create each bootstrap sample, we randomly selected diagnoses with replacement from the individual-level numerator data before aggregating and combining with the denominator data. The negative binomial GLM was modelled on these aggregated data. The frequency of selection for each variable was recorded as well as the final AIC of each model. The final model was chosen by establish a cutoff for frequency with which the model arose among the 1000 bootstraps. This reduced model was compared by mean square error (MSE) to a model without ecological covariates using one-third of the data reserved for testing purposes. Mean squared error (MSE) is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Where y_i is the observed value and $\hat{f}(x_i)$ is the predicted value for the i^{th}observation of n total observations [110].

Spatial Prediction with Stochastic Variable Selection

In the second approach, each diagnosis available in the individual-level data was randomly assigned to 1 of 10 partitions following a k-fold procedure [110]. Then, each combination of nine partitions was used to create a training dataset with the corresponding single partition, the “hold-out” fold, used to create a test dataset. The individual case data was aggregated and combined with population estimates to create strata-specific numerators, denominators and incidence rates by the following covariates: country of birth, age, sex, years in residence. These data were then used to create puma-specific observed cases and expected cases; the expected cases were calculated by standardizing observed cases by country of birth, age, sex, and years in residence. A similar operation was done with the corresponding single, “hold-out” partition, thus creating both training (9 parts in 10) and testing (1 part in 10) PUMA-specific data standardized to the aforementioned covariates.

Similar to prior work, the data were modelled using a spatial Poisson general linear mixed model BYM using the ecological variables listed above, though these data were aggregated to PUMA strata, instead of PUMA, age, sex, country of birth and years in residence as was done previously [101]. The spatial model described previously is commonly used in disease mapping applications, but here was extended with stochastic variable selection to assess the predictive utility of the ecological variables. Stochastic variable selection (SVS) applies a mixture model to each beta in the linear regression [106, 111].[19,25] Following the notation of Kleinschmidt et al. and Lunn et al., O_i is defined as the observed number of diagnoses occurring in area i; E_i is defined as the expected number of diagnoses for the same area [52, 86]. Additionally, we define $\eta_i \equiv E[O_i]$ and assume that $O_i \sim Poisson(\eta_i)$. The transformed linear regression is then:

$$log(\eta_i) = O_i + \alpha + \boldsymbol{\beta X} + \varphi_i \#(1)$$

where $\boldsymbol{\beta}$ denotes the vector of covariate betas, $\boldsymbol{X}$ denotes the vector of covariates and φ_i denotes a spatially-correlated random effects term defined by the following [85]:

$$\varphi_i | \varphi_{-i} = N\left(\overline{\varphi_\iota}, \frac{{\sigma_\varphi}^2}{n_i}\right)$$

$$\overline{\varphi_\iota} = \frac{1}{n_i} \sum_{j \in neighbors\ of\ i} \varphi_i$$

Neighbors of area i were defined with queen-style contiguity [87]. A map of Los Angeles County vintage 2000 PUMAs is shown in Figure 1.

To set up the mixture models for beta, we start with the prior probability of each individual possible model which is defined as:

$$P_w = \frac{1}{m}$$

where w is an index for each model and m is the number of models. The matrix M was constructed such that each possible combination of the putative covariates was represented as an individual row with 0 or 1 indicating the exclusion or inclusion of the particular variable in the specific model:

$$M = \begin{pmatrix} a_{11} & \cdots & a_{1j} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mj} \end{pmatrix}$$

where m is the number of models and j is number of putative covariates. In this case 14 variables were considered, resulting in 16,384 possible models and thus a 16,384 by 14 matrix. Covariate betas were defined by the following mixture model:

$$\beta_j \sim N(0, \tau_j)$$

$$\tau_j(\gamma_j) = \begin{cases} 0.1\gamma_j, & \gamma = 1 \\ 10\gamma_j, & \gamma = 0 \end{cases}$$

$$\gamma_j = M[k, j]$$

$$k_w \sim Categorical(P_w)$$

The indicator variable γ, which is confined to the interval (0,1), signifies the degree of support for the corresponding variable in the model. Indicator values near zero connote little support for the inclusion of the corresponding variable. The τ mixture essentially ensures that covariates with a probability close to 1 are included in the mode and those with probabilities close to zero are excluded. Additional priors were set as follows:

$$\alpha \sim U(-\infty, +\infty)$$

$$\varphi \sim G(0.5, 2000\ )$$

The model applied to the 10 training datasets. Each MCMC run had 2 chains of 100,000 iterations each with a burn-in period of 10,000. Thinning was not used as many no longer recommend the practice [112]. Model diagnostics were calculated including effective size. Support for individual variables was assessed by summing γ_j all iterations. Likewise, support for individual models was assessed by summing k_w across all iterations. Each resulting model was compared against the corresponding test dataset to compute an MSE and a k-fold cross validation metric was calculated using the following:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

R version 3.4, R Studio version 1.0.143 and variety of packages were used to manage and analyze data [51, 55, 57, 58, 60-63, 65, 88-90, 92, 93, 113-118]. Bayesian models were run in OpenBUGS version 3.2.2 rev 1012 [52].

Results

Non-spatial Models

Correlation plots and tree plots showed highly-correlated pairs among the 15 ecological variables chosen. The redundancy test produced a smaller group of 7 covariates: proportion receiving SNAP, proportion linguistically isolated, average number of rooms, proportion with incomplete plumbing, average number of units per structure and average income. The reduced set of ecological variables was examined with correlation and tree plots.

Results of the bootstrap process are detailed in tabular form. The frequency of selection in final model by variable for 1000 bootstrap replicates is shown in Table 6. Age, sex, country of birth and years in residence were all forced into the model based on performance in earlier studies and, as such, have frequency of selection of 1000 among 1000 bootstrap replicates [75]. The next most frequently selected variables were proportion in SNAP, average Federal poverty level, and proportion linguistically isolated all of which were selected for inclusion in more than 94% of the final models. In terms of the most frequently selected final models, the model including proportion on SNAP, proportion linguistically isolated, population density, average Federal poverty level, average family size, and proportion with a high education equivalent or higher was the final model among 6% of replicates. Table 7 includes frequencies of other final models.

Model specification, fit and predictive ability for the three non-spatial models are listed in Table 8. The minimum model, which was defined based on previous studies, is used as the baseline for comparison. The model constructed by naïve reduction had a moderately improved AIC in comparison to the baseline model. In addition, the model resulting for the bootstrap stepwise elimination method was a slight improvement in fit from the naïve reduction model. However, the predictive ability of these models was nearly identical. After running the models on the reserved test data, the mean squared error (MSE) for the minimum model was 0.21. There was negligible difference in the MSE between the baseline model and either the naïve reduction model or the bootstrap stepwise elimination model which yielded MSEs of 0.23 and 0.21 respectively.

Spatial Models

MCMC from all 10 folds had reasonable effective sizes for variables of interest (beta, gammas and k). Indicator variables from all proposed ecological variables across all 10 training partitions showed more support for average poverty and socio-economic index than other variables (Figure **5**). Among proposed models and across all 10 folds, support was highest for the model without any ecological variables henceforth called the "null" model (Figure ***6***). We ran the null model against the reserved test partitions and computed mean squared error for each fold and a cross validation metric (**Table 9**). MSEs between the non-spatial and spatial models are not directly comparable because of differences in included covariates.

DISCUSSION

TB-related ecological covariates at the PUMA level did not improve prediction of TB disease incidence rates in our models. In contrast to what has been observed in other locales, we did not observe correlation between TB incidence rates and TB-related ecological covariates specifically those which may serve as proxies for socio-economic status, crowding, population density and low access to health care [99, 100, 119-121].

In both non-spatial and spatial approaches, we were unable to substantially improve upon existing models with PUMA-level ecological variables [75, 101]. Results from non-spatial prediction showed limited benefit to adding the proposed ecological covariates. Similarly, results from spatial prediction models also showed limited benefit of ecological variables. Due to the preponderance of previous ecological and non-ecological literature showing the association between TB incidence and socio-economic variables, it is reasonable to consider that these ecological variables failed to improve the prediction of TB incidence because of the lack of specificity among ecological covariates at the PUMA level. PUMAs are designed to encompass approximately 100,000 persons; the amount of variation in socio-economic variables among those within a given PUMA is large, even in Los Angeles County where many PUMAs are geographically compact. It is also possible that country of birth and PUMA already capture much of the predictive information in the ecological variables tested.

There is some indication that ecological covariates at the census tract level can improve prediction of TB incidence [122]. Further research should include attempts using smaller geographic areas as well as attachment of ecological variables to strata other than geographies. For example, can assignment of ecological variables to country of birth strata improve the prediction of TB incidence?

While these avenues may incrementally improve prediction models, we would advocate for the collection of socio-economic data on TB cases, specifically the highest grade of education attained. Education level can be easily aligned with the American Community Survey and is recognized as a key component of socio-economic status (SES) [123]. Furthermore, national standards for TB reporting are revised decennially and new standards will be adopted in 2020. Currently, the standard TB surveillance has limited information of socio-economic status, though indication of employment and homeless are available. Also, a crude SES proxy could be created using the type of practice at which the case was first diagnosed.

By collecting socio-economic data at the case level using definitions harmonized to the American Community Survey, jurisdictions could calculate incidence rates by socio-economic status and adjust for SES in other analyses. These analyses have a significant advantage over ecological studies in that they are not subject to the ecological fallacy. In addition, improvement or changes to TB surveillance should work to include variables that directly align with those classifying the population-at-risk. Domestic TB control programs have increasingly drawn attention to the high proportion of cases (85%) estimated to be the result of reactivation of latent TB infection. Together with increased in interest in higher rates of progression to disease among persons with medical co-morbidities, such as diabetes, this further recommends socio-economic variables because of they can serve as proxies for access to care and quality of self-care.

These analyses have a number of limitations. The most prominent limitation is the lack of granularity among the available ecological data. Misclassification of key variables cannot be ruled out and using estimates from different data sources, though both mature, opens the door to errors due to the cross-misclassification of data. Case ascertainment and survey error remain important limitations and are discuss fully in prior work [75].

CONCLUSION

This study showed no improvement to TB disease incidence rate prediction models when including PUMA-level, socio-economic ecological variables. This was true both in non-spatial and spatial models and using multiple approaches to construct the prediction models including bootstrap backwards elimination, stochastic variable selection and k-fold partitioning. Socio-economic conditions are known to be important contributors to TB infection and progression to disease and these findings represent a failure to detect rather than a detection of absence. Improvement in prediction of TB incidence is sought in the expansion of routine surveillance to include collection SES data, especially given the importance of access to care among latent TB infected persons and the utility of SES data in predicting that access.

Table 6: Frequency of selection in final model by variable in 1000 bootstrap replicates

Variable	N
Age	1000
Country of birth	1000
Sex	1000
Years in residence	1000
Proportion in SNAP	994
Average Federal poverty level	957
Proportion linguistically isolated	944
Average family size	834
Population density	752
Proportion with high school education equivalent or higher	643
Average Hauser-Warren socio-economic index	486
Average income	432
Average socio-economic index	413
Proportion with incomplete plumbing	389
Proportion renting	354
Average number of rooms	232
Proportion with any health insurance	226
Average number of units in structure	210

Table 7: Heat map of Variable Inclusion and Frequency of final models in 1000 bootstrap replicates*

	Model															
Variables included in model	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Average Federal poverty level	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Proportion in SNAP	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Proportion linguistically isolated	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Average family size	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■
Population density	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■
Proportion with high school education eq. or higher	■		■	■	■	■	■	■	■	■	■		■	■	■	■
Average Hauser-Warren socio-economic index		■	■	■	■			■	■	■	■	■	■			■
Average socio-economic index		■		■	■			■		■		■				■
Proportion renting			■				■			■	■		■			
Proportion with incomplete plumbing						■					■		■	■		
Average income											■	■		■		
Proportion with any health insurance															■	■
Average number of rooms				■												
Average number of units in structure								■								
Frequency of model in 1000 bootstrap replicates	64	28	24	23	22	22	18	17	11	11	10	10	10	10	10	9
Relative Frequency	6%	3%	2%	2%	2%	2%	2%	2%	1%	1%	1%	1%	1%	1%	1%	1%

* models below 1% frequency not displayed

Table 8: Non-spatial models – specification, fit and predictive ability

Model	Variable†	AIC	MSE
Minimum	Country of birth, age, sex, years in residence	9206.38	0.2136
Naïve reduction	Minimum model plus: proportion on SNAP, proportion linguistically isolated, population density, proportion with incomplete plumbing, average number of units per structure, and average income	9011.44	0.2300
Bootstrap stepwise elimination	Minimum model plus: proportion on SNAP, proportion linguistically isolated, population density, average Federal poverty level, average family size, and proportion with a high education equivalent or higher	8990.00	0.2079

† *Variables shared by the naïve reduction model and the bootstrap stepwise elimination model are highlighted in grey*

Figure 5: Posterior Frequency of Variables by Fold

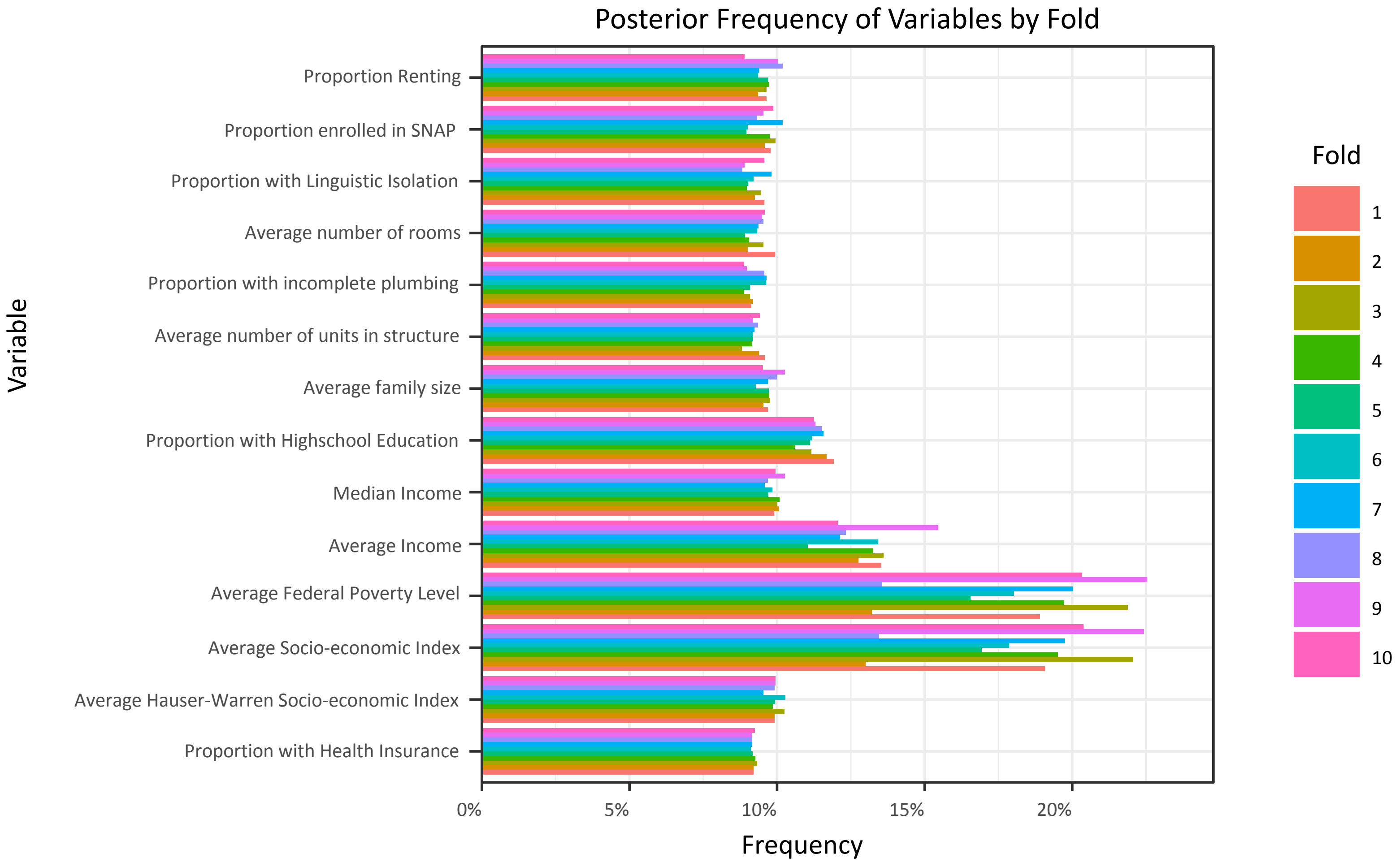


Figure 6: Posterior Proportion by Fold for top 10 Most Frequent Models across All Folds

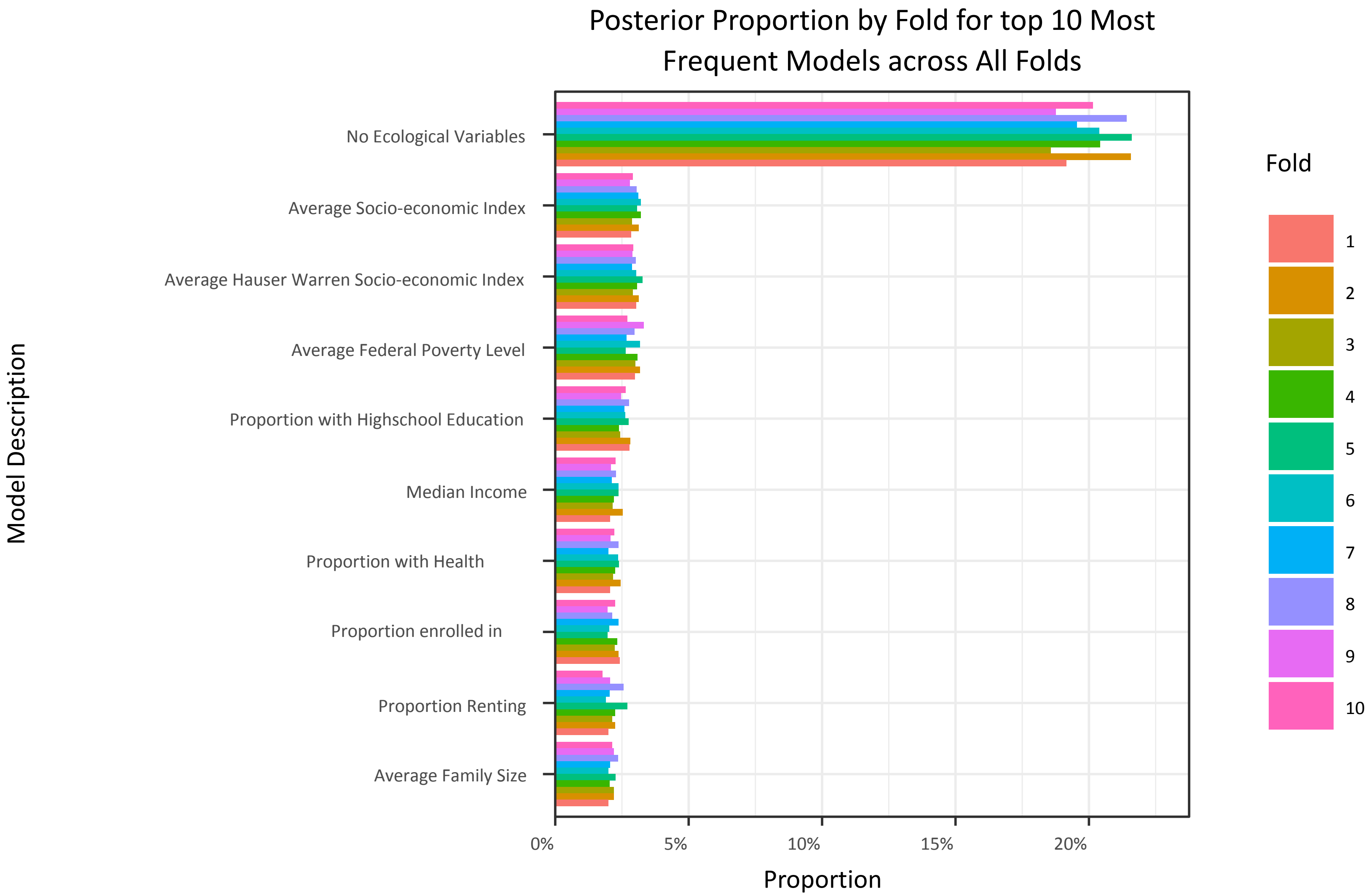


Table 9: Summary of Mean Square Error and Cross Validation Metric for Spatial Models

Fold	$(y_i - \hat{f}(x_i))^2$	MSE
1	436.468	0.693
2	343.154	0.545
3	270.435	0.429
4	479.420	0.761
5	438.125	0.695
6	334.141	0.530
7	439.771	0.698
8	389.847	0.619
9	343.040	0.545
10	327.570	0.520

	Sum of MSE	CV
All Folds	6.035	0.603

Conclusion

This thesis examined TB disease incidence among foreign-born persons in Los Angeles County, 2005-2011 using three analytical approaches. In the first approach, we examined TB incidence rates using non-spatial data only. The main conclusions of the non-spatial analysis are that: 1) TB disease incidence rates vary substantially by country of birth and therefore country-of-birth-specific incidence rates should be reported whenever possible; 2) TB disease incidence rates by country of birth can be calculated easily using available data from TB surveillance and the American Community Survey; 3) Poisson and Negative Binomial distributions are reasonable approximations of TB disease incidence provided that the proportion of recent transmission is low and there are sufficient covariates such that over-dispersion is limited; and 4) uncertainty in Census-derived population estimates used in the incidence rate calculations does not in this case affect point estimates and has limited effect on credibility intervals.

In the second approach, we examined TB incidence rates this time recognizing the spatial context of the disease. From the spatial analysis we conclude that: 1) TB disease incidence rates are heterogeneous across Los Angeles County even when accounting for important covariates such as age, sex, country of birth and years in residence; 2) unadjusted TB disease incidence rates estimation is limited by sparse data but this issue can be mitigated using spatial smoothing; 3) spatial models, building on non-spatial models from the first analytical approach, are reasonable approximation of this communicable disease.

In the third and final approach, we constructed prediction models building on non-spatial and spatial models. The main conclusion of this approach was that PUMA-level ecological variables, including poverty, linguistic isolation and educational attainment among others, did not improve the prediction over simpler models. Others have reported improvements in prediction using census tract ecological variables. It remains to be seen whether ecological variables can reliably improve prediction models for TB disease incidence rates.

In context, these results identify and fill a notable gap: country-of-birth-specific TB disease incidence rates are not regularly reported at the local, state or national level. We have shown here that calculating these rates is relatively straightforward and that necessary data for the denominator is accessible. Concerns about further modelling of TB disease incidence rates with standard distributions in count models are justified. However, we show empirically that standard distributions perform well in this context, presumably due to the low proportion of cases estimated to result from recent transmission. Together with careful construction of count models and diligent use of model diagnostics, we believe standard distributions can reliably be used and reasonable concerns regarding these methods can be addressed. We have also shown that we need not–and perhaps should not–be constrained to county geographies when analyzing TB incidence data even when including country of birth in the analysis. With this additional granularity, sparse data can easily limit analyses but with the application of spatial smoothing, we can extract useful information while addressing issues arising from sparse data. Finally, in an effort to improve the prediction of TB incidence rates, we have shown that PUMA-level ecological variables failed to improve both simple non-spatial models and more complex spatial models. While we still believe that ecological variables can improve prediction of TB incidence, we are forced by this result to consider repeating these efforts using alternate geographies and perhaps a wider range of ecological predictors.

Appendices

Appendix 1: Tuberculosis 2009 Case Definition – CSTE position statement 09-ID-65 [124, 125]

Tuberculosis (TB) (*Mycobacterium tuberculosis*)
2009 Case Definition

CSTE Position Statement(s)

- 09-ID-65

Clinical Description

A chronic bacterial infection caused by *Mycobacterium tuberculosis*, usually characterized pathologically by the formation of granulomas. The most common site of infection is the lung, but other organs may be involved.

Clinical Criteria

A case that meets all the following criteria:

- A positive tuberculin skin test or positive interferon gamma release assay for *M. tuberculosis*
- Other signs and symptoms compatible with tuberculosis (TB) (e.g., abnormal chest radiograph, abnormal chest computerized tomography scan or other chest imaging study, or clinical evidence of current disease)
- Treatment with two or more anti-TB medications
- A completed diagnostic evaluation

Laboratory Criteria for Diagnosis

- Isolation of *M. tuberculosis* from a clinical specimen,* **OR**
- Demonstration of *M. tuberculosis* complex from a clinical specimen by nucleic acid amplification test,** **OR**
- Demonstration of acid-fast bacilli in a clinical specimen when a culture has not been or cannot be obtained or is falsely negative or contaminated.

Case Classification

Confirmed

A case that meets the clinical case definition or is laboratory confirmed

Comments

A case should not be counted twice within any consecutive 12-month period. However, a case occurring in a patient who had previously had verified TB disease should be reported and counted again if more than 12 months have elapsed since the patient completed therapy. A case should also be reported and counted again if the patient was lost to supervision for greater than 12 months and TB disease can be verified again. Mycobacterial diseases other than those caused by *M. tuberculosis* complex should not be counted in tuberculosis morbidity statistics unless there is concurrent tuberculosis.

*Use of rapid identification techniques for *M. tuberculosis* (e.g., DNA probes and mycolic acid high-pressure liquid chromatography performed on a culture from a clinical specimen) are acceptable under this criterion.

** Nucleic acid amplification (NAA) tests must be accompanied by culture for mycobacteria species for clinical purposes. A culture isolate of *M. tuberculosis* complex is required for complete drug susceptibility testing and also genotyping. However, for surveillance purposes, CDC will accept results obtained from NAA tests approved by the Food and Drug Administration (FDA) and used according to the approved product labeling on the package insert, or a test produced and validated in accordance with applicable FDA and Clinical Laboratory Improvement Amendments (CLIA) regulations.

Appendix 2: Quick guide to population estimates by country of birth

1) Navigate to American Factfinder Advanced Search, https://factfinder.census.gov
2) Enter "B05006" into table name box
3) Select "B05006: PLACE OF BIRTH FOR THE FOREIGN-BORN POPULATION" from available selections
4) Enter jurisdiction for which you have case counts by country of birth
5) Select the standardized jurisdiction name from selection
6) From resulting table list, select appropriate year/table
    a. For one year of case counts, choose the appropriate year and the dataset marked "ACS 1-year estimates", e.g. "2015 ACS 1-year estimates"
    b. Choose the appropriate year and dataset marked "ACS 5-year estimates." ACS 5 year estimates are labelled with the final year of data collection. For example, the 2015 5-year estimates represented data from 2011-2015 averaged. Equivalent to one year estimate but with smaller MOE
7) Download data

Appendix 3: OpenBUGS Model code

```
model{
  for (i in 1:numrec) {

    numer[i] ~ dpois(mu[i]) #Poisson numerator declaration

    # alternatively negative binomial numerator declaration:
    # numer[i] ~ dnegbin(p[i], r)
    # p[i] <- r/(r + mu[i])



    log(mu[i]) <-
      alpha +
      beta.cob[cobn[i]] +
      beta.age[agecatn[i]] +
      beta.sex[sexn[i]] +
      beta.year[yearn[i]] +
      beta.yres[yres_catn[i]] +
      log(denom[i])
  }


  alpha ~ dnorm(0,0.00001)

  beta.cob[1] <- 0
  for (j in 2:numcob) {
    beta.cob[j] ~ dnorm(0,0.001)
  }

  beta.age[1] <- 0
  for (k in 2:numage) {
    beta.age[k] ~ dnorm(0,0.001)
  }

  beta.sex[1] <- 0
  for (m in 2:numsex) {
    beta.sex[m] ~ dnorm(0,0.001)
  }

  beta.year[1] <- 0
  for (n in 2:numyear) {
    beta.year[n] ~ dnorm(0,0.001)
  }
  beta.yres[1] <- 0
  for (p in 2:numyres) {
    beta.yres[p] ~ dnorm(0,0.001)
  }
}
```

References

1. Centers for Disease Control and Prevention, *Reported Tuberculosis in the United States, 2015*.
2. World Health Organization, *Global Tuberculosis Report*. 2016.
3. TB Alliance. *TB is now the World's Leading Infectious Killer*. 9/24/2017]; Available from: https://www.tballiance.org/news/tb-now-worlds-leading-infectious-killer.
4. Harvard University Library. *Contagion:Historical Views of Disease and Epidemics: Tuberculosis in Europe and North America, 1800-1922.* 2014 9/15/2014]; Available from: http://ocp.hul.harvard.edu/contagion/tuberculosis.html.
5. World Health Organization, *Bugs, Drugs & Smoke: Stories from public health.*, ed. W.H. Organization. 2011, Geneva, Switzerland.
6. Pastor, M., et al., *California Immigrant Integration Scorecard - Technical Report*. 2012.
7. Centers for Disease Control and Prevention (CDC). *TB Basic Facts*. 9/23/2017]; Available from: https://www.cdc.gov/tb/topic/basics/default.htmhttps://www.cdc.gov/tb/topic/basics/default.htmhttps://www.cdc.gov/tb/topic/basics/default.htm.
8. Centers for Disease Control and Prevention (CDC), *Core Curriculum on Tuberculosis: What the Clinician Should Know.* 2013.
9. Comstock, G.W., *Untreated inactive pulmonary tuberculosis. Risk of reactivation.* Public Health Rep., 1962. **77**: p. 461-470.
10. Getahun, H., R.E. Chaisson, and M. Raviglione, *Latent Mycobacterium tuberculosis Infection.* N Engl J Med, 2015. **373**(12): p. 1179-80.
11. Sterling, T.R., et al., *Three months of rifapentine and isoniazid for latent tuberculosis infection.* N. Engl. J. Med., 2011. **365**(23): p. 2155-2166.
12. Centers for Disease Control and Prevention (CDC). *Latent Tuberculosis Infection: A Guide for Primary Health Care Providers*. 9/24/2017]; Available from: https://www.cdc.gov/tb/publications/ltbi/targetedtesting.htm.
13. World Health Organization (WHO), *Global Health Observatory data repository*. 2015.
14. Centers for Disease Control and Prevention (CDC), *Reported Tuberculosis in the United States, 2015*. 2016.
15. Centers for Disease Control and Prevention (CDC), *Reported Tubeculosis in the United States*. 2012.
16. Los Angeles County Department of Public Health Tuberculosis Control Program, *TB Cases by Country of Origin, Los Angeles County, 2007-2011 - Table 7.* 2012, Los Angeles County Department of Public Health, Tubculosis Control Program: Los Angeles.
17. Los Angeles County Department of Public Health Tuberculosis Control Program, *Tuberculosis Cases by Foreign/U.S.-born, Los Angeles County, 2007-2011*. 2012: Los Angeles.
18. Centers for Law and the Public's Health, *Tuberculosis Control Laws and Policies: A Handbook for Public Health and Legal Practitioners*. 2009.
19. Thacker, S.B., J.R. Qualters, and L.M. Lee, *Public health surveillance in the United States: evolution and challenges.* MMWR Suppl, 2012. **61**(3): p. 3-9.

20. California Health & Safety Code § 121361
21. California Health & Safety Code § 121362
22. U.S. Department of Labor, *Occupational Exposure to Tuberculosis; Proposed Rule; Termination of Rulemaking Respiratory Protection for M. Tuberculosis; Final Rule; Revocation*, in *29*, U.S. Federal Register, Editor. 2003.
23. U.S. Department of Labor Office of Safety and Health Administration. *Tuberculosis*. 9/23/2017]; Available from: https://www.osha.gov/SLTC/tuberculosis/.
24. California Code of Regulations § 5199
25. Taylor, Z., et al., *Controlling tuberculosis in the United States. Recommendations from the American Thoracic Society, CDC, and the Infectious Diseases Society of America.* MMWR Recomm. Rep., 2005. **54**(RR-12): p. 1-81.
26. Ruggles, S., et al., *Integrated Public Use Microdata Series: Version 6.0* 2015, University of Minneapolis: Minneapolis.
27. U. S. Census Bureau, *QuickFacts: Los Angeles County, California*. 2016.
28. U.S. Census Bureau, *Annual Estimates of the Resident Populations*.
29. Los Angeles County Chief Executive Office, *Cities within the County of Los Angeles*.
30. U.S. Census Bureau, *Growth in Urban Population Outpaces Rest of Nation*. 2012.
31. Los Angeles Times. *Koreatown*. Mapping L.A. 9/24/2017]; Available from: http://maps.latimes.com/neighborhoods/neighborhood/koreatown/.
32. Los Angeles Times. *Westlake*. Mapping L.A. 9/24/2017]; Available from: http://maps.latimes.com/neighborhoods/neighborhood/westlake/.
33. Census Reporter. 9/24/2016].
34. U.S. Census Bureau, *American Community Survey: Table B05006*.
35. Los Angeles County Department of Public Health. *About Us*. 9/24/2017]; Available from: http://publichealth.lacounty.gov/phcommon/public/aboutus/aboutdisplay.cfm?ou=ph&prog=ph&unit=ph.
36. Centers for Disease Control and Prevention (CDC), *Reported Tuberculosis in the United States, 2015*. US Department of Health and Human Services.
37. Public Health England, *Tuberculosis in England, 2016 Report*.
38. U.S. Census Bureau, *American Community Survey, 2015 American Community Survey 5-Year Estimates, Table B05002*. 2015.
39. Zuber, P.L., et al., *Tuberculosis among foreign-born persons in Los Angeles County, 1992-1994.* Tuber. Lung Dis., 1996. **77**(6): p. 524-530.
40. Cain, K.P., et al., *Tuberculosis among foreign-born persons in the United States.* JAMA, 2008. **300**(4): p. 405-412.
41. Centers for Disease Control and Prevention (CDC), *Report of Verified Case of Tuberculosis (RVCT) Manual.* 2009.
42. U.S. Census Bureau, *A Compass for Understanding and Using American Community Survey Data: What PUMS Data Users Need to Know*. 2009: Washington, DC.
43. U.S. Census Bureau, *American Community Survey Design and Methodology*. 2014.
44. U.S. Census Bureau, *American Community Survey Design and Methodology: Chapter 12 - Variance Estimation*.
45. IPUMS USA. *Replicate Weights in the American Community Survey*. Available from: https://usa.ipums.org/usa/repwt.shtml.

46. World Health Organization, *TB drug resistance types.* 2015.
47. Centers for Disease Control and Prevention (CDC). *Extensively Drug-Resistant Tuberculosis (XDR TB)*. 9/24/2017]; Available from: https://www.cdc.gov/tb/publications/factsheets/drtb/xdrtb.htm.
48. New York State Department of Health. *Rates Based on Small Numbers - Statistics Teaching Tools*. 4/28/2017]; Available from: https://www.health.ny.gov/diseases/chronic/ratesmall.htm.
49. Baker, B.J., C.D. Jeffries, and P.K. Moonan, *Decline in Tuberculosis among Mexico-Born Persons in the United States, 2000–2010.* Ann. Am. Thorac. Soc., 2014. **11**(4): p. 480-488.
50. Hilbe, J.M., *Modeling Count Data*. 2009.
51. R Development Core Team, *R: A language and environment for statistical computing*. 2016, R Foundation for Statistical Computing: Vienna, Austria.
52. Lunn, D., et al., *The BUGS project: Evolution, critique and future directions.* Statistics in Medicine, 2009. **28**(25): p. 3049-3067.
53. NIMBLE Development Team, *NIMBLE: An R Package for Programming BUGS models*. 2017.
54. California Department of Public Health TB Control Branch, *Report on Tuberculosis in California, 2015,*. 2016.
55. RStudio Team, *RStudio: Integrated Development Environment for R*. 2016.
56. Grolemund, G. and H. Wickham, *Dates and Times Made Easy with lubridate.* Journal of Statistical Software, 2011. **40**(3): p. 1-25.
57. Hadley Wickham, R.F., Lionel Henry, Kirill Müller, *dplyr: A Grammar of Data Manipulation*. 2017.
58. Harrell Jr, F.E. and contributions from Charles Dupont and many others, *Hmisc: Harrell Miscellaneous*. 2017.
59. Hilbe, J.M., *COUNT: Functions, Data and Code for Count Data*. 2016.
60. Matt Dowle, A.S., *data.table: Extension of `data.frame`*. 2017.
61. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2009: Springer-Verlag New York.
62. Wickham, H., *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. 2017.
63. Wickham, K.M.a.H., *tibble: Simple Data Frames*. 2017.
64. Lumley, T., *Analysis of Complex Survey Samples.* Journal of Statistical Software, 2004. **9**(1): p. 1-19.
65. Gelman, S.S.a.U.L.a.A., *R2WinBUGS: A Package for Running WinBUGS from R.* Journal of Statistical Software, 2005. **12**(3): p. 1-16.
66. Washington State Department of Health, *Guidelines for Working with Small Numbers*.
67. Centers for Disease Control and Prevention (CDC). *NAMCS/NHAMCS - Reliability of Estimates*. Available from: https://www.cdc.gov/nchs/ahcd/ahcd_estimation_reliability.htm.
68. Zuber, P.L., et al., *Long-term risk of tuberculosis among foreign-born persons in the United States.* JAMA, 1997. **278**(4): p. 304-307.
69. Walter, N.D., et al., *Persistent latent tuberculosis reactivation risk in United States immigrants.* Am. J. Respir. Crit. Care Med., 2014. **189**(1): p. 88-95.

70. California Department of Public Health TB Control Branch, *California Tuberculosis Risk Assessment*. 2016.

71. France, A.M., et al., *A field-validated approach using surveillance and genotyping data to estimate tuberculosis attributable to recent transmission in the United States.* Am J Epidemiol, 2015. **182**(9): p. 799-807.

72. Curtis, A.B.M., E.; McKenna, M.; Onorato, I. M., *Completeness and timeliness of Tuberculosis Case Reporting: A Multistate Study.* American Journal of Preventative Medicine, 2001. **20**(2): p. 108-112.

73. Winston, C.A., et al., *Unexpected decline in tuberculosis cases coincident with economic recession - United States, 2009.* BMC Public Health, 2011. **11**: p. 846.

74. Centers for Disease Control and Prevention (CDC), *Trends in Tuberculosis — United States, 2010.* Morbidity and Mortality Weekly Report, 2011.

75. Readhead, A., *Tuberculosis Incidence Estimation among Foreign-born Persons, Los Angeles County 2005-2011*. 2017.

76. Los Angeles County Department of Public Health Tuberculosis Control Program, *Tuberculosis in Los Angeles County: Surveillance Report 2013*. 2015: Los Angeles, CA.

77. Feske, M.L., et al., *Including the third dimension: a spatial analysis of TB cases in Houston Harris County.* Tuberculosis, 2011. **91 Suppl 1**: p. S24-33.

78. Agarwal, S., et al., *Spatial-temporal distribution of genotyped tuberculosis cases in a county with active transmission.* BMC Infect. Dis., 2017. **17**(1): p. 378.

79. Oppong, J.R., et al., *Foreign-Born Status and Geographic Patterns of Tuberculosis Genotypes in Tarrant County, Texas.* Prof. Geogr., 2007. **59**(4): p. 478-491.

80. U.S. Preventive Services Task Force (USPSTF), *Draft Recommendation Statement: Latent Tuberculosis Infection: Screening*.

81. Zuber, P.L.F.K., L. S.; Binkin, N. J.; Tipple, M. A.; Davidson, P. T. , *Tuberculosis among foreign-born persons in Los Angeles County, 1992-1994.* Tubercle and Lung Disease, 1996. **77**: p. 524-530.

82. U.S. Census Bureau, *Public Use Microdata Areas (PUMAs).* 2012.

83. Le Polain De Waroux, O., et al., *The epidemiology of gonorrhoea in London: a Bayesian spatial modelling approach.* Epidemiol. Infect., 2014. **142**(1): p. 211-220.

84. Souza, W.V., et al., *Tuberculosis in intra-urban settings: a Bayesian approach.* Trop. Med. Int. Health, 2007. **12**(3): p. 323-330.

85. Besag, J., J. York, and A. Mollié, *Bayesian image restoration, with two applications in spatial statistics.* Ann. Inst. Stat. Math., 1991. **43**(1): p. 1-20.

86. Kleinschmidt, I., et al., *Rise in malaria incidence rates in South Africa: a small-area spatial analysis of variation in time trends.* Am. J. Epidemiol., 2002. **155**(3): p. 257-264.

87. Bivand, R., E.J. Pebesma, and V. Gómez-Rubio, *Applied spatial data analysis with R*, in *Use R!* 2013, Springer,: New York, NY. p. 1 online resource.

88. Bivand, R., J. Hauke, and T. Kossowski, *Computing the Jacobian in Gaussian spatial autoregressive models: An illustrated comparison of available methods.* Geographical Analysis, 2013. **45**(2): p. 150-179.

89. Bivand, R., T. Keitt, and B. Rowlingson, *rgdal: Bindings for the Geospatial Data Abstraction Library*. 2017.

90. Bivand, R. and N. Lewin-Koh, *maptools: Tools for Reading and Handling Spatial Objects*. 2017.

91. Bivand, R. and G. Piras, *Comparing Implementations of Estimation Methods for Spatial Econometrics.* Journal of Statistical Software, 2015. **63**(18): p. 1-36.

92. Bivand, R. and C. Rundel, *rgeos: Interface to Geometry Engine - Open Source (GEOS)*. 2017.

93. Richard A. Becker; Thomas P Minka; Allan R. Wilks, R.B.A.D., *maps: Draw Geographical Maps*. 2017.

94. Centers for Disease Control and Prevention (CDC), *CDC WONDER Online Database National Tuberculosis Surveillance System,*, Online Tuberculosis Information System (OTIS), Editor. 2015.

95. Cain, K.P., et al., *Tuberculosis Among Foreign-Born Persons in the United States.* Journal of the American Medical Association, 2008. **300**(4): p. 405-412.

96. Alex J. Goodell , P.B.S., Rick Vreman , John Metcalfe , Travis C. Porco , Pennan M. Barry , Jennifer Flood , Suzanne Marks , Adithya Cattamanchi , Jim G. Kahn ,, *Prospects for Elimination: An Individual-based Model to Assess TB Control Strategies in California.*, in *American Thoracic Society*. 2016: San Francisco, CA.

97. Ozcaglar, C., et al., *Epidemiological models of Mycobacterium tuberculosis complex infections.* Math Biosci, 2012. **236**(2): p. 77-96.

98. Harling, G. and M.C. Castro, *A spatial analysis of social and economic determinants of tuberculosis in Brazil.* Health Place, 2014. **25**: p. 56-67.

99. Wubuli, A., et al., *Socio-Demographic Predictors and Distribution of Pulmonary Tuberculosis (TB) in Xinjiang, China: A Spatial Analysis.* PLoS One, 2015. **10**(12): p. e0144010.

100. Hargreaves, J.R., et al., *The social determinants of tuberculosis: from evidence to action.* Am. J. Public Health, 2011. **101**(4): p. 654-662.

101. Readhead, A., *Spatial Distribution of TB incidence, Los Angeles County 2005-2011*. 2017.

102. Santa Clara County Public Health Department, T.P.a.C.P., *TB Fact Sheet*. 2017.

103. Myers, W.P., et al., *An Ecological Study of Tuberculosis Transmission in California.* Am J Public Health, 2006. **96**(4): p. 685-90.

104. Austin, P.C. and J.V. Tu, *Bootstrap Methods for Developing Predictive Models.* Am. Stat., 2004. **58**(2): p. 131-137.

105. Harrell Jr., F.E., *Regression modeling strategies*. 2001, Nashville: Springer.

106. O'Hara, R.B. and M.J. Sillanpää, *A review of Bayesian variable selection methods: what, how and which.* Bayesian Anal., 2009. **4**(1): p. 85-117.

107. Harrell Jr, F.E., *Hmisc: Harrell Miscellaneous*. 2017.

108. Wei, T. and V. Simko, *corrplot: Visualization of a Correlation Matrix*. 2016.

109. Venables, W.N. and B.D. Ripley, *Modern Applied Statistics with S*. 2002: Springer.

110. James, G.W., D.; Hastie T; Tibshirani, R;, *An Introduction to Statistical Learning - with Applications in R | Gareth James | Springer*. 2017.

111. Reich, B.B.J. and S. Ghosh, *A Review of Bayesian Variable Selection.*

112. Link, W.A. and M.J. Eaton, *On thinning of chains in MCMC.* Methods Ecol. Evol., 2012. **3**(1): p. 112-115.

113. Franziska Hoffgaard and with contributions from Philipp Weil and Kay Hamacher, *BioPhysConnectoR*. 2013.

114. Wickham, H., *Reshaping Data with reshape Package.* Journal of Statistical Software, 2007. **21**(12): p. 1-20.

115. Curtis, S.M., *mcmcplots: Create Plots from MCMC Output*. 2015.

116. de Valpine, P., et al., *nimble: Flexible BUGS-Compatible System for Hierarchical Statistical Modeling and Algorithm Development*. 2017.

117. Lumley, T., *survey: analysis of complex survey samples.* Journal of Statistical Software, 2016. **9**(1): p. 1-19.

118. Bivand, R., *classInt: Choose Univariate Class Intervals*. 2017.

119. Maciel, E.L.N., et al., *Spatial patterns of pulmonary tuberculosis incidence and their relationship to socio-economic status in Vitoria, Brazil.* Int. J. Tuberc. Lung Dis., 2010. **14**(11): p. 1395-1402.

120. Barbosa, D.S., et al., *Spatial analysis for identification of priority areas for surveillance and control in a visceral leishmaniasis endemic area in Brazil.* Acta Trop., 2014. **131**: p. 56-62.

121. Ojo, O.B., S. Lougue, and W.A. Woldegerima, *Bayesian generalized linear mixed modeling of Tuberculosis using informative priors.* PLoS One, 2017. **12**(3): p. e0172580.

122. Barry, P., *TB incidence Ecological Analysis by Census Tract*. 2017.

123. Oakes, J.M. *Measuring Socioeconomic Status*. 9/24/2017]; Available from: http://www.esourceresearch.org/

124. Council of State and Territorial Epidemiologists (CSTE), *Public Health Reporting and National Notifcation for Tuberculosis*. 2009.

125. Centers for Disease Control and Prevention (CDC). *Tuberculosis (TB) (Mycobacterium tuberculosis) 2009 Case Definition*. 2009 10/7/2017]; Available from: https://wwwn.cdc.gov/nndss/conditions/tuberculosis/case-definition/2009/.