JOURNAL OF
EMERGING INVESTIGATORS

Article

A five-year retrospective analysis of Tuberculosis risk factors and their variability in the United States

Yash Kini[1]*, Sophia Diaz Gaviria[2]*, Edgar A. Diaz[3,4], Soumya Kini[5]
[1] James Madison High School, Vienna, Virginia
[2] West Hills High School, Santee, California
[3] Department of Research & Health Promotion, San Ysidro Health, San Diego, California
[4] Alzheimer Gut Microbiome Project, Duke University, Durham, North Carolina
[5] Health Strategy, Operations and Systems, MITRE Corporation, McLean, Virginia

*These authors contributed equally to this work

SUMMARY

Tuberculosis (TB) persists as a major health concern globally and disproportionately affects non-U.S.-born individuals and certain racial/ethnic groups in the United States. We hypothesized that disparities in TB rates among different racial/ethnic groups in the United States are influenced by a complex interaction of several risk factors including socioeconomic conditions, cultural practices, environmental exposures, and access to public health services. Using the Centers for Disease Control and Prevention Online Tuberculosis Information System database, we analyzed cases and incidence rates of TB from 2018 to 2022, evaluating disparities by demographic and geographic factors and identifying key risk predictors through machine learning models. We also compared this data to the preceding five years (2013-2017) to understand the variability in risk factors over time. Our results revealed that non-U.S.-born individuals accounted for 71.8% of TB cases and experienced much higher incidence rates compared to U.S.-born populations. Racial/ethnic differences were also pronounced, with the highest incidence rates in Asian and Pacific Islanders, while age-related risk varied between U.S.-born and non-U.S.-born groups. Our feature importance analyses identified origin of birth, older age, and male sex as the strongest predictors of TB risk, with variations in predictor significance over time, reflecting evolving TB epidemiology. These findings underscore the need for targeted interventions to reduce TB disparities and improve outcomes in high-risk populations. By addressing the complex relationship of socioeconomic factors, healthcare access, migration, and public health priorities, this research contributes to our understanding of the TB landscape in the United States.

INTRODUCTION

Tuberculosis (TB), a contagious airborne bacterial infection primarily affecting the lungs, is caused by *Mycobacterium tuberculosis (M. tb)*. It remains a significant global public health issue – approximately a quarter of the world population is infected with latent TB. In 2023 alone, 10.8 million people developed active TB, consistently making TB one of the leading causes of death by a single infectious agent (1-3). While briefly surpassed by COVID-19 in 2020, TB's impact on global mortality remains substantial (4, 5).

TB infection exists in two states: latent and active. In latent TB, individuals carry the bacterium asymptomatically and cannot transmit the disease (1). However, if their immune system weakens, latent TB can progress to active TB, leading to serious health complications (1). Active TB is characterized by respiratory and systemic symptoms such as cough, fever, weight loss, and fatigue. Individuals with active TB are highly infectious and can transmit the disease through prolonged close contact, generally after eight hours (6).

Despite having one of the world's lowest TB incidence rates (2.9 cases per 100,000 persons), the United States (U.S.) is currently experiencing a resurgence in TB cases, posing a growing public health concern, particularly among vulnerable populations (7, 8). This resurgence underscores the persistent threat of TB, even in low-incidence settings, and emphasizes the importance of ongoing vigilance and targeted interventions. The epidemiological landscape of TB in the U.S. is complex and multifaceted, influenced by both domestic and global factors. A notable proportion of TB cases over a five-year period between 2018 and 2022 – 6,148 (73.8%) – were reported in non-U.S.-born individuals, predominantly from Asian countries with high TB prevalence (9). This increased rate of TB in non-U.S.-born individuals highlights the role of migration in TB transmission dynamics (9). In many low-incidence nations, a substantial number of TB diagnoses occur among individuals born in higher-prevalence regions, often resulting from the reactivation of latent TB infections acquired prior to their migration (10, 11). While TB is most frequently observed in individuals aged 65 and older, children aged 5-14 years experienced the most significant increase in cases in 2023, with an alarming 45% rise (6). Geographically, the states experiencing the highest incidence of TB include Alaska, Hawaii, California, New York, and Texas, suggesting regional disparities in TB burden and transmission patterns (8).

TB control and elimination depend on early detection, prompt treatment, and prevention of secondary cases. However, addressing TB requires more than just medical interventions. Social determinants of health significantly influence all aspects of TB, from exposure risk to treatment outcomes (12). The current treatment regimen for drug-sensitive TB is lengthy and complex, often leading to challenges in adherence. When combined with poverty and limited healthcare access, this contributes to the emergence of drug-resistant strains, which are both more complex and costly to treat (1). Individual beliefs, stigma, and misconceptions about TB can deter individuals from seeking

https://doi.org/10.59720/25-027

care. These barriers, financial constraints, and limited access to healthcare facilities aggravate the problem. Socioeconomic factors such as poverty, inadequate housing, malnutrition, substance abuse, and comorbidities like HIV/AIDS further increase vulnerability to TB and hinder access to diagnosis and treatment (13).

While these factors are widely recognized, there remains a critical need to better understand the complex ways in which social determinants, race/ethnicity, and migration interact to shape TB susceptibility, particularly in a country like the U.S. with a complex demographic landscape and a history of health disparities. We examined this gap by analyzing data from the Centers for Disease Control and Prevention (CDC) from 2018 to 2022 (the most recent five years of available data at the time of this study) to investigate the complex interaction of social determinants, race/ethnicity, and migration in shaping TB susceptibility within the U.S. (14). This period was significantly impacted by the COVID-19 pandemic and its associated changes in social dynamics, healthcare access, and public health priorities. These factors likely influenced TB transmission and reporting, contributing to an initial decline in cases in 2020 during the pandemic, potentially due to increased social isolation and reduced healthcare utilization. However, this trend has reversed in 2021, with an upward trend in newly diagnosed cases by the end of 2022, which may be attributed to a backlog of diagnoses delayed by COVID-19 disruptions (15). This post-pandemic increase, with case numbers reaching levels last seen in 2013, is occurring across all age groups and most states, with a disproportionate 18% rise among non-U.S.-born individuals (16). Despite a low overall TB rate in the U.S., persistent disparities remain, particularly among foreign-born individuals and those identifying as Native Hawaiian/Pacific Islander, American Indian/Alaska Native, or Black (7).

Considering these observations, we hypothesized that disparities in TB rates among different racial/ethnic groups in the U.S. are influenced by a complex interaction of risk factors, including socioeconomic conditions, cultural practices, environmental exposures, and access to public health services. By exploring the trends in TB rates over five-years in the U.S., dataset, we aim to understand the factors contributing to disparities in TB rates across different populations and inform targeted interventions to address this pressing public health concern.

RESULTS

TB cases and incidence rates in the U.S.

We started our analysis by looking at trends in TB cases and incidence rates in the U.S. from 2018 to 2022 to help understand the annual trends in TB cases to inform public health strategies (**Figure 1**). During this time period, the number of reported TB cases generally decreased, with notable changes corresponding to the World Health Organization's defined COVID-19 pandemic period of March 2020 to May 2023 (16). TB cases declined from 8,997 in 2018 to 7,171 in 2020, representing a 20.3% reduction. However, as the Food and Drug Administration began approving COVID-19 vaccines, cases rebounded to 8,331 by 2022. Similarly, TB incidence rates (in cases per 100,000) followed a downward trend, declining from 2.8 in 2018 to 2.2 in 2020 before increasing to 2.5 in 2022.

We further examined TB proportions and incidence rates by origin of birth (U.S.-born and non-U.S.-born individuals) in the U.S. from 2018 to 2022 to identify variability in TB cases between these populations (**Figure 2**). A total of 41,264 TB cases were reported during this time-period. Non-U.S.-born individuals accounted for 71.8% (29,637 cases) of all reported TB cases, with an average incidence rate of 13.1 cases per 100,000 population – substantially higher than the corresponding rate of 0.8 per 100,000 (or 27.8%; 11,486 of

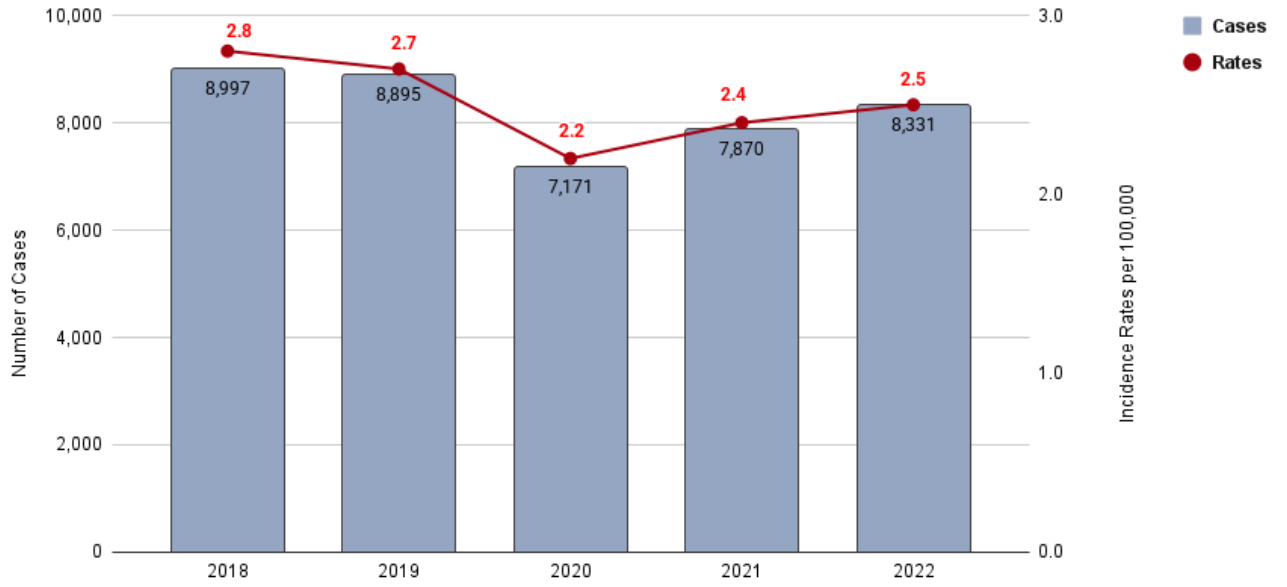

Figure 1. Tuberculosis cases in the United States from 2018 to 2022. The blue vertical bars represent Tuberculosis (TB) cases by year. The red line represents TB incidence rates per 100,000 individuals for the same years and are plotted on a secondary right axis. Data from Centers for Disease Control and Prevention (CDC) Online Tuberculosis Information System (OTIS) repository (14).

https://doi.org/10.59720/25-027

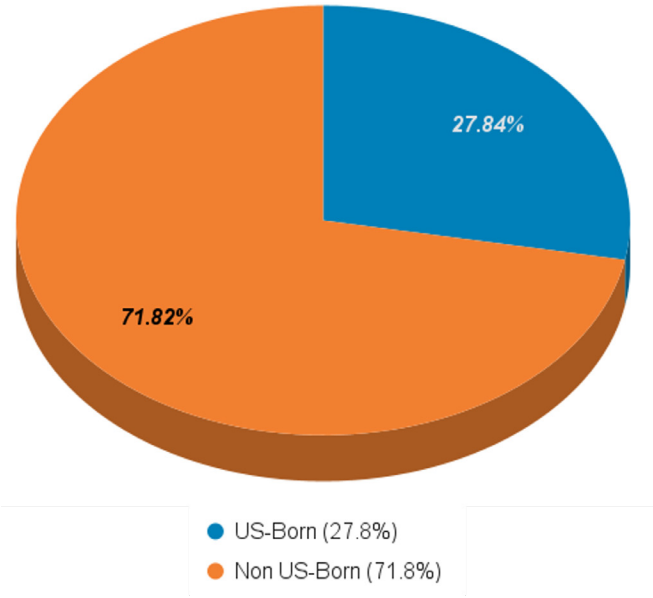

Figure 2. Percentages of U.S. and non-U.S.-born individuals with TB from 2018 to 2022. Percentage of Tuberculosis (TB) cases reported among non-U.S.-born individuals (orange) and U.S.-born (blue) out of 41,264 total cases.

the total cases) among U.S.-born individuals.

Risk factors for TB in the U.S.

Differences in TB cases and incidence rates by race/ethnicity in the U.S. from 2018 to 2022 were particularly notable for non-U.S.-born Asian, Native Hawaiian, or Other Pacific Islander group which had the highest TB incidence rates when compared to other race/ethnicity categories (**Figure 3**). Among Asian, Native Hawaiian, or Other Pacific Islander individuals, 94% (14,408 cases) of reported TB were among non-U.S.-born individuals, corresponding to the highest incidence rate of 24.3 cases per 100,000 population. Hispanic or Latino individuals had the second-highest incidence rate of 7.0 per 100,000, with 78% of cases (9,896 cases) occurring among non-U.S.-born individuals.

Black or African American individuals showed an incidence rate of 5.9 per 100,000, with 48% of cases (3,668 cases) involving non-U.S.-born persons. In contrast, White individuals demonstrated the lowest incidence rate of 0.8 per 100,000, with 27% of cases (1,255 cases) occurring among non-U.S.-born individuals. Although American Indian or Alaska Native populations accounted for only 1% of all cases, they had a relatively high incidence rate of 5.6 per 100,000.

Prior studies have identified age as a critical TB risk factor; therefore, we next examined whether age patterns differed among the high-incidence racial and ethnic groups. The distribution of TB cases by age group among U.S.-born and non-U.S.-born individuals across three racial and ethnic groups that had the highest incidence rates (Asians, Black/African Americans, and Hispanics/Latinos) revealed notable differences that are described in the paragraph below in age-related TB risk between these populations and between U.S.-born and non-U.S.-born individuals (**Figure 4**).

For U.S.-born individuals (total cases: 7,545), those under 24 years of age constituted the highest TB cases for Hispanic/Latino and Asian individuals when compared to other racial/ethnic groups (**Figure 4A**). Conversely, those between 45–64 years represented the highest TB cases for Black/African Americans, showing that middle-aged individuals carried a notable burden in this group. TB cases among individuals aged 65 years and older and those in the 25–44 years range were lower overall across all the U.S.-born populations.

The trends differed notably for non-U.S.-born individuals (total cases: 27,942). Among Hispanics/Latinos and Black/African Americans, those between 25–44 years of age had the highest number of TB cases (**Figure 4B**). For Asians, however, those above 65 years of age were the highest, contrasting trends seen in the U.S.-born populations.

Feature importance model results

To better understand which factors influence TB incidence the most we developed a feature importance ranking model. We developed feature importance rankings from the models for evaluating TB risk factors using two datasets: 2013-2017 and 2018-2022 (**Figure 5A-B**).

In the 2013-2017 dataset, the most influential features were race being Asian or Pacific Islander, 65 or older age group and males (**Figure 5A**). All three features were positively associated with TB risk. These findings suggested that older Asian or Pacific Islander men experienced the highest TB burden during this period. Other important factors included origin of birth (not U.S.-born, >50% group) and HIV-positive groups with 4-9% HIV-positive rates, reinforcing the impact of origin of birth and HIV status on TB vulnerability. Conversely, features such as under 24 years of age, race being White and female demonstrated negative importance, indicating a lower TB risk associated with these groups.

The 2018-2022 dataset demonstrated similar trends, with origin of birth (not U.S.-born, >50% group) emerging as the top predictor, followed by 65 or older age group and male (**Figure 5B**). Notably, the importance of race being Asian or Pacific Islander remained high, while HIV-related features continued to play a crucial role. The reduced importance of under 24 years of age and race being White further confirmed the persistent lower TB risk in these groups during this period of study.

TB risk factors by state

We used a Tableau dashboard to visualize TB incidence rates across racial and ethnic groups in the U.S. from 2018-2022 (**Figure 6A-D**) (17). The dashboard was developed for this study on the Tableau Public server (18). The dashboard highlights geographic and demographic disparities in TB cases, allowing for a nuanced understanding of how factors such as origin of birth, race/ethnicity, age, and sex contribute to TB distribution. We identified a few key notable insights from the dashboard.

First, for American Indian or Alaska Native individuals, TB incidence rates reached as high as 65.52 per 100,000, with the highest burden concentrated in Alaska (**Figure 6A**). Compared to other racial/ethnic groups, TB cases among American Indian or Alaska Native individuals were geographically concentrated.

For Asian or Pacific Islander individuals, TB incidence rates were as high as 51.24 per 100,000, clustering in states like California, Texas, and New York (**Figure 6B**). TB incidence rates for Black or African American individuals reached a maximum of 75.97 per 100,000, with the highest burden concentrated in the northeastern U.S. and urban areas

https://doi.org/10.59720/25-027

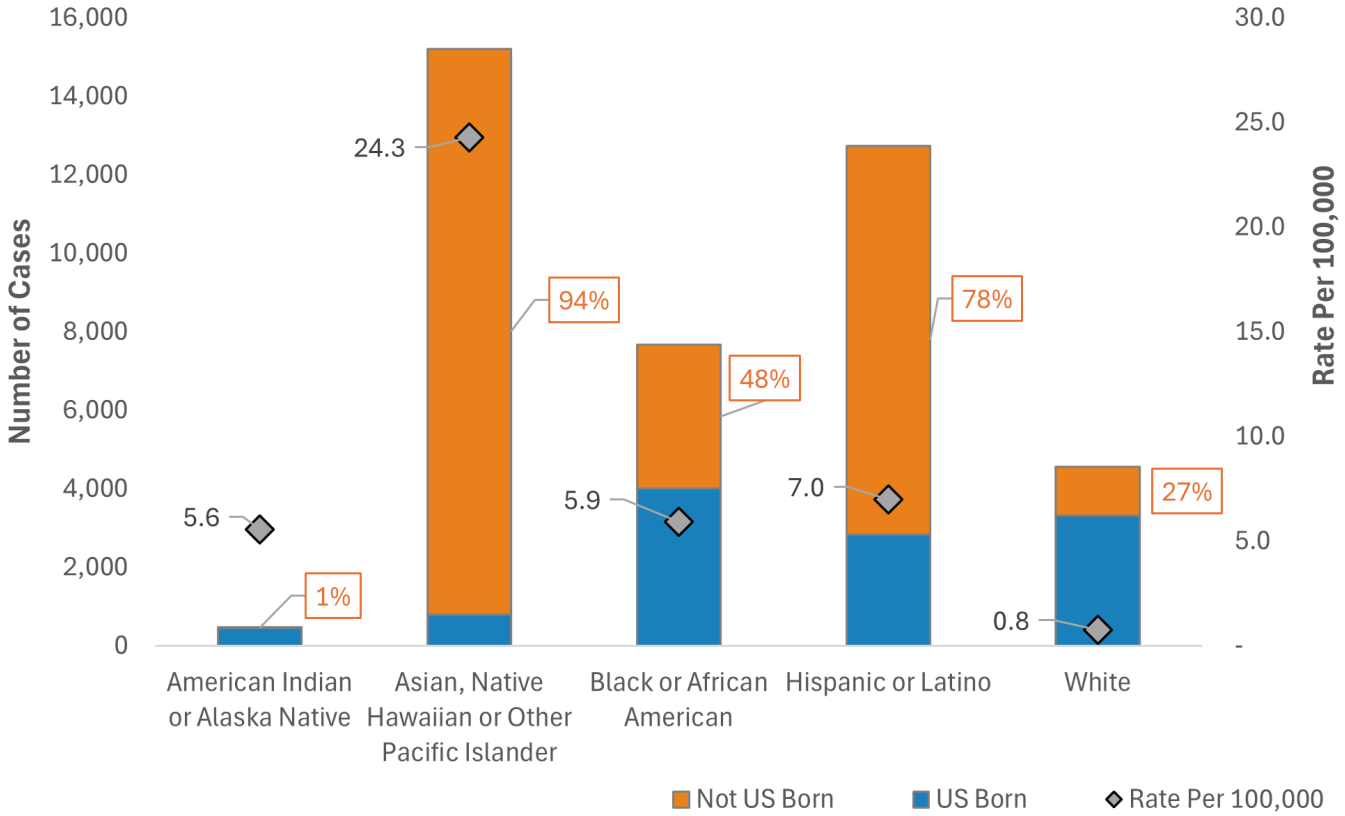

Figure 3. Tuberculosis Cases and Incidence Rates by Race/Ethnicity from 2018 to 2022. The orange section of the vertical bars represent the number of Tuberculosis (TB) cases reported among non-U.S.-born individuals for a particular racial/ethnic group. Numbers in the square boxes show the percentage of cases represented by U.S.-born individuals for a particular racial/ethnic group. The blue sections of the vertical bars represent the number of TB cases reported among U.S.-born individuals for a particular racial/ethnic group. The grey diamonds and corresponding numbers represent TB incidence rates per 100,000 individuals for racial/ethnic group and are plotted on a secondary right axis.

along the Midwest and East Coast, the most potent cluster being located in Maine (**Figure 6C**). Finally, for Hispanic or Latino individuals, TB incidence rates reached a maximum of 16.31 per 100,000, with clustering in the southeastern U.S., particularly in Tennessee, Georgia, Arkansas, and Louisiana (**Figure 6D**).

DISCUSSION

Our findings reveal striking disparities in TB rates across origin of birth, race/ethnicity, and age groups, aligning with global trends that indicate an uneven decline in TB incidence and mortality rates. Building on this overview, this study provides a comprehensive analysis of TB incidence in the U.S. from 2018 to 2022, a period marked by the impact of the COVID-19 pandemic (19). TB trends during this time period indicated that various pandemic-related factors influenced TB incidence and reporting. The stark difference in TB incidence rates between U.S.-born and non-U.S.-born individuals underscored the disproportionate burden of TB among non-U.S.-born populations. While public health measures such as lockdowns and social distancing may have initially reduced opportunities for TB transmission, these same restrictions and the general disruption to healthcare services likely led to delayed diagnosis and treatment initiation for existing TB cases, potentially allowing for continued community transmission (19, 20). Furthermore, the overwhelming focus on COVID-19 and similarities in symptoms might have also contributed to an underestimation of TB incidence due to underreporting or misdiagnosis.

Our analysis supports our hypothesis that notable disparities in TB susceptibility exist within the U.S. population, primarily driven by an individual's place of birth, and that these disparities interact with socioeconomic factors and access to healthcare. Individuals born outside the U.S. experience a disproportionate burden of TB, with considerably higher incidence rates than their U.S.-born counterparts. This disparity likely reflects increased prior exposure to TB in high-prevalence countries, reactivation of latent TB infection, barriers in accessing healthcare, and adherence to prescribed treatment plans in the U.S (21). These findings stress the need for tailored interventions, including enhanced screening, accessible treatment programs, and community health outreach initiatives designed explicitly for non-U.S.-born populations (22). This need is clearly illustrated in states like California and Hawaii, where a majority of TB cases occur in foreign-born populations. In California, the disparity in TB rates between foreign-born individuals (16.6 per 100,000) and U.S.-born persons (1.3 per 100,000) is particularly striking. Similarly, in Hawaii, 86% of individuals with TB in 2023 were non-U.S.-born, as reported by the Hawaii Department of Health (23). These states highlight the importance of addressing the unique challenges faced by immigrant communities in accessing TB prevention and care services.The feature importance analysis further emphasized

https://doi.org/10.59720/25-027

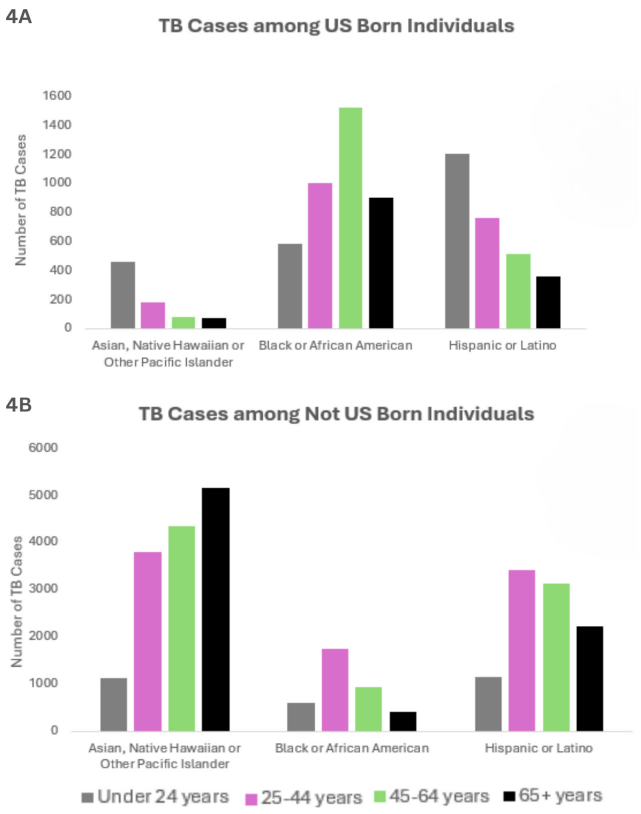

Figure 4. Age as a risk factor from 2018 to 2022. Tuberculosis (TB) cases among A) U.S.-born and B) non-U.S.-born individuals for the three racial/ethnic groups that had the highest incidence rates: Asians, Black/African Americans, and Hispanics/Latinos. Each ethnic group is further broken down by age.

the complex relationship of factors influencing TB risk. Origin of birth, older age, and male sex consistently emerged as strong predictors of TB incidence, further confirming the importance of origin of birth highlighted above and reinforcing the need to prioritize these high-risk groups in public health interventions. The evolving importance of features like HIV positivity across the two timeframes examined (2013-2017 and 2018-2022) suggests dynamic shifts in TB epidemiology and highlights the need for ongoing surveillance and adaptive strategies.

Beyond the origin of birth, our analysis further highlights the critical role of age in TB risk, with distinct patterns observed across different demographics. The high TB burden among U.S.-born Hispanics/Latinos under 24 years of age may reflect socioeconomic vulnerabilities and challenges in accessing preventative care. On the other hand, the predominance of cases in older age groups among non-U.S.-born individuals could indicate the reactivation of latent TB infection due to waning immunity or the presence of comorbidities. These age-specific trends underscore the importance of tailoring public health strategies to the unique needs of each demographic. For instance, youth-focused educational programs, community-based screening initiatives, and accessible healthcare services for younger individuals could help mitigate TB risk in this population. Similarly, targeted interventions for older adults, such as latent TB infection testing and treatment, may help prevent disease reactivation and transmission.

The diverse challenges different states face highlights the need for public health interventions tailored to specific populations and their unique needs. In Alaska, the highest TB prevalence nationwide is concentrated among Alaska Natives, particularly in the mountainous southwestern region, thus demonstrating the impact of geography and healthcare access. Indeed, this cluster drives the national TB incidence rate observed among American Indian/Alaska Native populations. Distinct challenges are also evident in Minnesota, with elevated TB rates among the Native Hawaiian and Pacific Islander population. Furthermore, our state-level analysis revealed a concentration of high TB rates among Black/African Americans in the northeastern U.S., likely reflecting the persistent impact of socioeconomic disparities and historical inequities in this region. Addressing these underlying social determinants of health is crucial to achieving TB elimination in this population. While our current data cannot confirm specific contributing factors intrinsic to the particular populations, observed disparities may be related to socioeconomic inequality, cultural practices and preferences, underlying health conditions, and the impact of COVID-19. These cases exemplify the need for public health interventions tailored to the unique challenges faced by diverse communities.

Our findings also highlight the impact of the COVID-19 pandemic on TB incidence. Our analysis revealed a 20.3% decline in reported cases between 2018 and 2020, potentially reflecting healthcare access and diagnosis disruptions, as well as increased social isolation and the adoption of preventive measures such as masking and improved ventilation and air filtering to reduce COVID-19 transmission. However, a concerning resurgence in cases was observed in the post-pandemic period, with case numbers rebounding to similar levels to those in 2013 (16). This underscores the importance of strengthening public health infrastructure and ensuring continued access to TB services, even during global health emergencies. The observed shift in the top predictor of TB risk from race/ethnicity to origin of birth between 2013-2017 and 2018-2022 may reflect changes in immigration and travel patterns, particularly in the post-pandemic era. This demonstrates the need for ongoing surveillance and adaptive strategies to address the evolving demographics of TB risk.

The findings of this study underscore the urgent need for comprehensive and multifaceted approaches to TB control in the U.S. Future research should prioritize investigating the social determinants of health that contribute to TB disparities, including poverty, housing insecurity, and limited access to healthcare. Rigorous evaluation of targeted interventions, such as community health worker programs, culturally tailored education campaigns, and enhanced screening initiatives, is crucial to identify effective strategies for reducing TB transmission and improving health outcomes in vulnerable populations (22). We also need to further investigate the long-term impact of the COVID-19 pandemic on TB control efforts to inform public health preparedness and ensure continued progress toward TB elimination. Policymakers and public health professionals must prioritize implementing evidence-based strategies to address TB disparities and protect the health of all communities. This includes increased investment in public health infrastructure, expanding access to affordable healthcare, and developing innovative TB prevention and

https://doi.org/10.59720/25-027

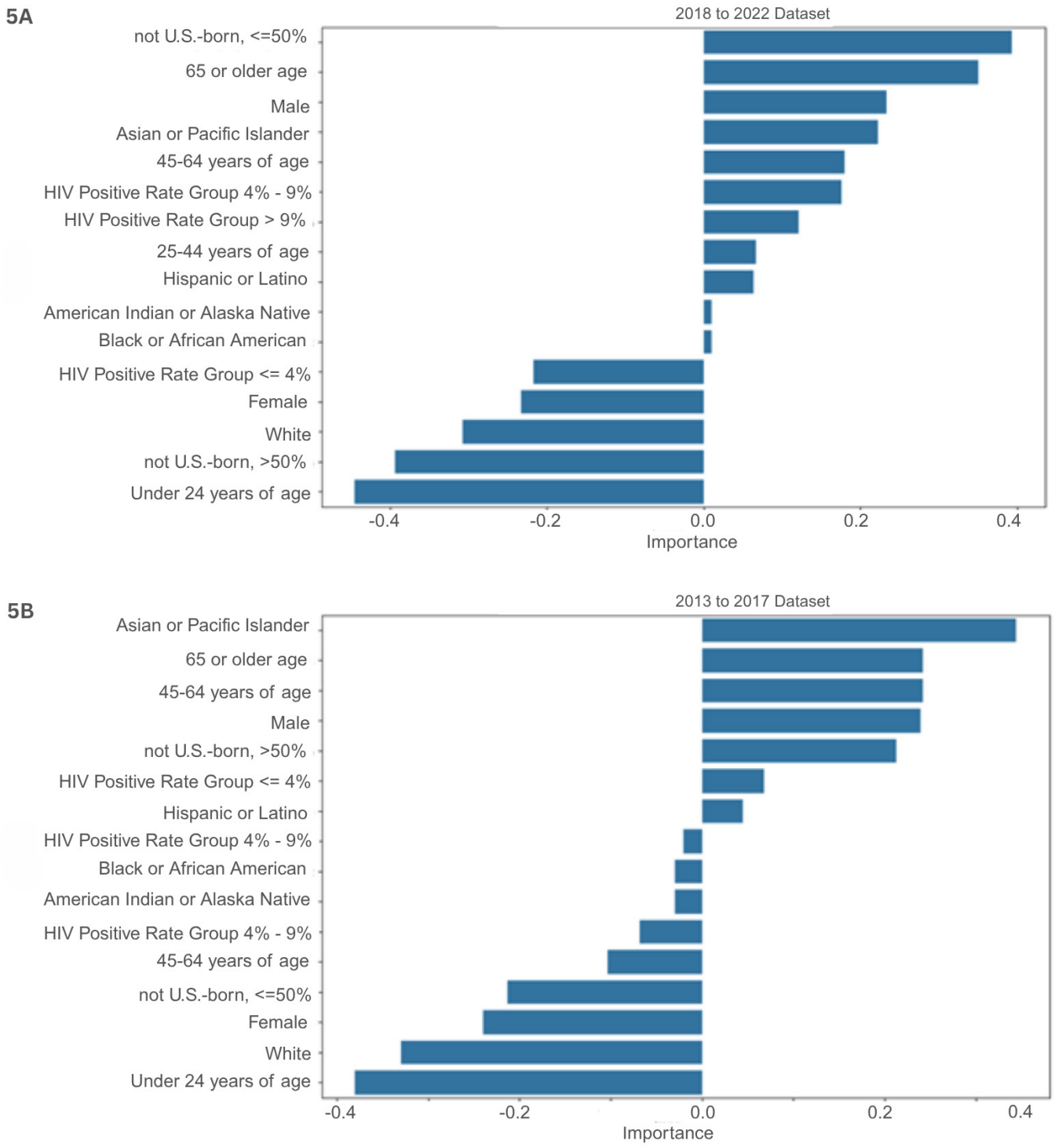

Figure 5. Feature importance of Tuberculosis risk factors. Risk factors for the A) 2013-2017 and B) 2018-2022 datasets. The contribution of each predictor to the model's predictions. A positive importance value indicates a positive correlation between the predictor and target variables; a negative importance value indicates a negative correlation between the predictor and target variables; an importance value of zero indicates no effect between the predictor and target variables.

care approaches. By working collaboratively and addressing the complex factors contributing to TB vulnerability, we can make significant progress in mitigating this persistent public health threat.

There are several limitations inherent to using secondary data sources. Firstly, the granularity of the available data was limited due to privacy concerns, necessitating the analysis of five years of aggregated information as a block. This precluded a more detailed examination of trends by state, including changes in TB incidence by age or ethnicity over time, and limited the inclusion of a robust set of risk factors. Secondly, as this study is observational, we can

https://doi.org/10.59720/25-027

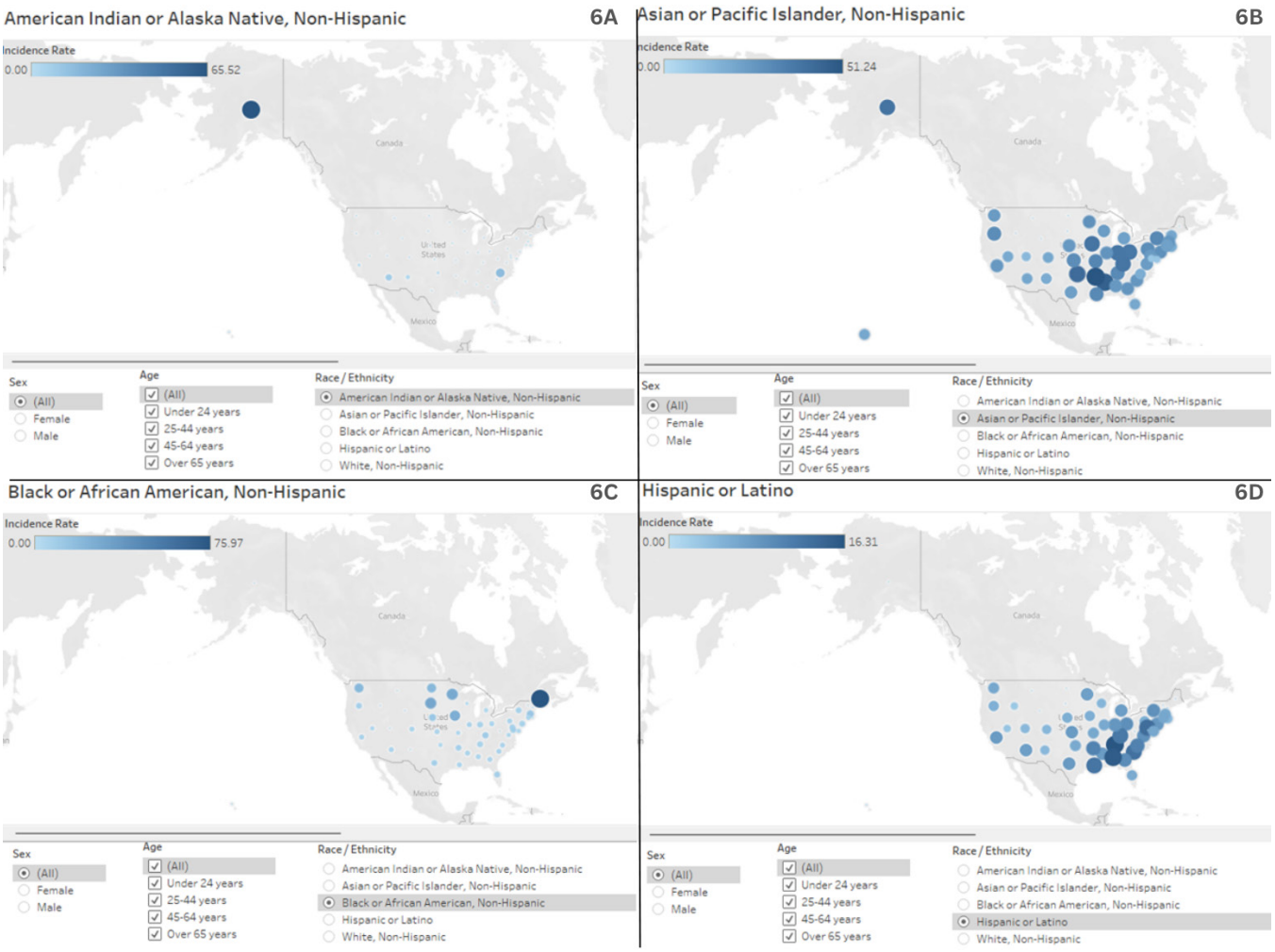

Figure 6. Tableau dashboard showing Tuberculosis rates by state from 2018 to 2022. Rates among A) American Indian/Alaska Native, B) Asian/Pacific Islander, C) Black/African American and D) Hispanic/Latino. Note the variable incidence rate scale on each chart, which is generated by Tableau based on the incidence rate range for the race group of interest selected.

only infer associations between variables and not causality. Additionally, it is possible that TB cases were underreported, particularly during the COVID-19 pandemic, due to limited access to healthcare and potential misattribution of deaths. Furthermore, the pandemic's impact on social interactions and heightened precautions likely influenced TB transmission dynamics, making it difficult to disentangle these effects from other factors. Finally, some of the reported cases in the post-COVID-19 period may represent older, previously undiagnosed cases rather than an actual increase in incidence. However, the continued upward trend suggests a genuine resurgence of TB. Limitations related to the Online Tuberculosis Information System (OTIS) (15) data itself should also be acknowledged. While OTIS is a valuable resource for TB surveillance, it is subject to potential biases in reporting practices across states and may not capture all cases of TB due to underreporting or misdiagnosis. At the time of this analysis, CDC's OTIS database had data published until 2022. The codebase developed for this study is available on the GitHub repository and can be re-run with more recently published data to investigate any changes.

MATERIALS AND METHODS

Data sources

To ensure the reliability and relevance of data for this study on TB cases, incidence rates, and associated risk factors, an extensive literature review and database search was conducted. The primary goal was to identify comprehensive data sources encompassing key variables such as TB cases, incidence rates, and sociodemographic risk factors. Searches were performed on PubMed Central, individual state health department websites, and the CDC repository. While multiple sources were examined, the CDC's OTIS emerged as the most robust and consistent dataset for the analysis (14). OTIS was ultimately chosen due to its comprehensive coverage of TB records from 1993 onward, available variables, and its consistent annual reporting.

The variables used in this study from the OTIS database included: i) State (all 50 U.S. states, the District of Columbia, and Puerto Rico), ii) Cases (the number of reported TB cases in each state annually), iii) Year (divided into two time-frames: 2013 to 2017 and 2018 to 2022 to evaluate temporal changes in TB trends), iv) Age groups (the patient's age group at the

https://doi.org/10.59720/25-027

time their case was initially reported to the health department as a suspected TB case), v) Sex (the biological sex of groups diagnosed with TB, categorized as male or female), vi) Bridged race/ethnicity (combined race and ethnic origin to reflect patients' self-identified racial and ethnic categories, as characterized by CDC), vii) Origin of birth (U.S.-born and non-U.S.-born individuals to assess location of birth as a risk factor), viii) HIV status (proportion of TB cases with known HIV infection, reflecting the interaction between TB and HIV).

Data preparation

The data collected from the OTIS database were systematically documented and organized using Google Sheets. Separate datasets were created for the 2013-2017 and 2018-2022 time periods to facilitate temporal comparisons. All variables used in the analysis were categorical. Additional columns were created, as needed, to categorize variables that were continuous in the OTIS dataset. For example, variables such as TB incidence rates and percentages of HIV-positive status were converted into categorical ranges to align with the analytical framework. This categorization ensured consistency in the data, accuracy in the models, and simplified the modeling process. An overview of the final variables and their respective categorical values is presented in **Table 1**.

Model development

The model development was conducted using JupyterLab, an interactive development environment widely used for data analysis and machine learning tasks (24). The predictor variables included age group, sex, race/ethnicity, origin of birth, and HIV status, representing the input features used to predict the target variable of TB incidence rate.

Feature importance charts were created to evaluate the relative influence of predictor variables on the target variable. These charts display the contribution of each predictor to the model's predictions, helping to identify the most pertinent factors driving TB outcomes. In a feature importance chart, a positive value indicates a positive correlation between the predictor and target variables; a negative value indicates a negative correlation between the predictor and target variables; a value of zero indicates no effect between the predictor and target variables. Seaborn, matplotlib, and scikit-learn Python packages were used for data conversions, visualizations of feature importance, and for investigating which variables had the most notable impact on TB incidence and prevalence. The complete codebase is available on GitHub (25).

TB dashboard

An interactive TB dashboard was developed using Tableau

Variable Name	Variables Values
Race	• American Indian or Alaska Native, Non-Hispanic • Asian or Pacific Islander, Non-Hispanic • Black or African American, Non-Hispanic • Hispanic or Latino • White, Non-Hispanic
Age	• Under 24 years of age • 25-44 years of age • 45-64 years of age • 65+ years of age
Sex	• Male • Female
Percentage of Non-U.S.-Born Individuals in a Group	• <=50% (Groups where the majority are U.S.-born) • >50% (Groups where the majority are not U.S.-born)
Percentage of HIV-Positive Individuals in a Group	• Groups with <4% HIV-positive rates • Groups with 4%-9% HIV-positive rates • Groups with >9% HIV-positive rates
Incidence Rate	• <=8 • Between 8 and 16 • >16

Table 1: Variables and their values used in the analyses. All variables used in the analysis were categorical and were converted into categorical ranges, as needed, to align with the analytical framework. This categorization ensured consistency in the data, accuracy in the models, and simplified the modeling process.

https://doi.org/10.59720/25-027

for visualizing TB incidence rates across various demographic groups and geographic locations (18). The dashboard utilized publicly available state-level TB data for the years 2018-2022, providing insights into TB incidence rates by sex, age group, and race/ethnicity.

ACKNOWLEDGMENTS

The authors wish to acknowledge Dave Duluk, data scientist at the MITRE Corporation for his guidance on the development and interpretation of the feature importance charts used to identify the most important predictors of TB risk.

Received: December 14, 2024
Accepted: February 14, 2026
Published: March 14, 2026

REFERENCES

1. Alsayed, Shahinda S., and Hendra Gunosewoyo. "Tuberculosis: Pathogenesis, current treatment regimens and New Drug Targets." *International Journal of Molecular Sciences*, vol. 24, no. 6, 8 Mar. 2023, p. 5202. https://doi.org/10.3390/ijms24065202.
2. Kwan, C. K., and J. D. Ernst. "HIV and Tuberculosis: A Deadly Human Syndemic." *Clinical Microbiology Reviews*, vol. 24, no. 2, 1 Apr. 2011, pp. 351–376. https://doi.org/10.1128/cmr.00042-10.
3. "Tuberculosis (TB)." World Health Organization, World Health Organization, www.who.int/news-room/fact-sheets/detail/tuberculosis. Accessed 27 May 2025.
4. Ledesma, Jorge R, et al. "Global, Regional, and National Age-Specific Progress towards the 2020 Milestones of the WHO End TB Strategy: A Systematic Analysis for the Global Burden of Disease Study 2021." *The Lancet Infectious Diseases*, vol. 24, no. 7, 1 Mar. 2024, pp. 698 – 725. https://doi.org/10.1016/s1473-3099(24)00007-0.
5. Baker, Rachel E., et al. "Infectious Disease in an Era of Global Change." *Nature Reviews Microbiology*, vol. 20, no. 20, 13 Oct. 2021, pp. 193 – 205. https://doi.org/10.1038/s41579-021-00639-z.
6. "Tuberculosis (TB)." *Centers for Disease Control and Prevention*. www.cdc.gov/tb. Accessed 10 Dec. 2024.
7. "Reported Tuberculosis in the United States, 2023." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, www.cdc.gov/tb-surveillance-report-2023/index.html. Accessed 27 May 2025.
8. Williams, Paula M., et al. "Tuberculosis — United States, 2023." MMWR. Morbidity and Mortality Weekly Report, vol. 73, no. 12, 28 Mar. 2024, pp. 265–270, https://doi.org/10.15585/mmwr.mm7312a4.
9. "Demographics." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 15 Nov. 2023, www.cdc.gov/tb/statistics/reports/2022/demographics.htm.
10. Talbot, Elizabeth A. "Tuberculosis among foreign-born persons in the United States, 1993-1998." *JAMA*, vol. 284, no. 22, 13 Dec. 2000, p. 2894, https://doi.org/10.1001/jama.284.22.2894.
11. Talwar, Amish, et al. "Association between birth region and time to tuberculosis diagnosis among non–US-born persons in the United States." *Emerging Infectious Diseases*, vol. 27, no. 6, June 2021, pp. 1645–1653, https://doi.org/10.3201/eid2706.203663.
12. Duarte, R., et al. "Tuberculosis, Social Determinants and Co-Morbidities (Including HIV)." *Pulmonology*, vol. 24, no. 2, Mar. 2018, pp. 115–119. https://doi.org/10.1016/j.rppnen.2017.11.003.
13. Lee, Ju-Yeun, et al. "Inadequate Housing and Pulmonary Tuberculosis: A Systematic Review." *BMC Public Health*, vol. 22, no. 1, 30 Mar. 2022, p. 622. https://doi.org/10.1186/s12889-022-12879-6.
14. "OTIS TB Data 1993-2022, Archive Request." *Centers for Disease Control and Prevention*. wonder.cdc.gov/tb-v2022.html. Accessed 27 Nov. 2024.
15. "Evidence, Policy, Impact. WHO Guide for Evidence-Informed Decision-Making." *World Health Organization*. https://www.who.int/publications/i/item/9789240039872. Accessed 10 Dec. 2024.
16. "Coronavirus Disease (Covid-19) Pandemic." World Health Organization, World Health Organization, www.who.int/europe/emergencies/situations/covid-19. Accessed 27 May 2025.
17. Batt, Steven, et al. "Learning Tableau: A Data Visualization Tool." *The Journal of Economic Education*, vol. 51, no. 3-4, 20 Aug. 2020, pp. 317-28, https://doi.org/10.1080/00220485.2020.1804503.
18. Kini, Yash. "Tuberculosis Cases and Incidence Rates by US States - 2018 to 2022." *Tableau Public*, Published 5 Dec. 2024, https://public.tableau.com/app/profile/yash.kini/viz/TuberculosisCasesandIncidenceRatesbyUSStates-2018to2022/Dashboard1. Accessed 18 August 2025.
19. Oh, Kyung Hyun, et al. "Changing Epidemic of Tuberculosis Amidst the Covid-19 Pandemic in the Western Pacific Region: Analysis of Tuberculosis Case Notifications and Treatment Outcomes from 2015 to 2022." *The Lancet Regional Health Western Pacific*, vol. 47, 5 Jun. 2024. https://doi.org/10.1016/j.lanwpc.2024.101104.
20. McQuaid, C. F., et al. "The impact of covid-19 on TB: A review of the Data." The International Journal of Tuberculosis and Lung Disease, vol. 25, no. 6, 1 June 2021, pp. 436–446, https://doi.org/10.5588/ijtld.21.0148.
21. Spence, Brigitte C, et al. "Patient-Reported Barriers to Treatment Initiation and Completion for Latent Tuberculosis Infection among Patients within a Large Integrated Health Care System in Southern California." *Journal of Public Health Management and Practice*, vol. 29, no. 3, 1 Mar. 2023, pp. 345-352, https://doi.org/10.1097/PHH.0000000000001711.
22. Joo, J. Y., and M. F. Liu. "Culturally Tailored Interventions for Ethnic Minorities: A Scoping Review." *Nursing Open*, vol. 8, no. 5, 9 Dec. 2020, pp. 2078-2090. https://doi.org/10.1002/nop2.733.
23. "Ka 'Oihana Olakino." *State of Hawaii, Department of Health*. https://health.hawaii.gov/tb/. Accessed 10 Dec. 2024.
24. "Jupyter." *Project Jupyter*. https://jupyter.org/. Accessed 15 Jan. 2025.
25. Kini, Yash. "Code and Data for the JEI Paper on Tuberculosis." *GitHub*, Published 8 Dec. 2024, https://github.com/yashkini1/TuberculosisJEI. Accessed 18 August 2025

https://doi.org/10.59720/25-027