

**Department of Epidemiology and Healthcare**

A Quantitative Analysis of the Factors Associated with Tuberculosis Pathology in Lima, Peru

Candidate Number: JYDW1

Student Number: 21081749

2024

A white card with black text

Description automatically generated

281

379

291

5,427

# Acknowledgements

My greatest thanks to my dissertation supervisor, who has guided me throughout this project. He has been extremely helpful in choosing a research topic, connecting me with project collaborators, and contributing valuable insight to my primary draft. I would also like to thank my supervisor’s PhD student for her consistent support and advice. This has been an exhilarating journey that has grounded my passion for research writing. I look forward to continuing my work with them in the near future and contributing to TB research in the field under their supervision.

I would also like to thank the course leads, professors, and my fellow coursemates, who have made my past three years at UCL a life-changing and unforgettable experience. Through my courses and curriculum, I have discovered the field I am truly passionate about, and for that, I am forever grateful.

In this section, I also acknowledge the use of artificial intelligence tools in the production of this report. ChatGPT4 (https://chat.openai.com/) was used to support the testing of my code written for the regression models and assist with code in data cleaning.

# Reflections

While writing this paper, I was able to apply the skills learned over the last three years of my degree in Population Health and Data Science in courses that ranged from quantitative analysis and immunology to the social determinants of health. I utilised comprehensive research skills that followed a systematic and scientific methodology and carried out complex data manipulation, statistical analysis, and visualisation on a secondary dataset. By focusing on tuberculosis, I have applied my understanding of infectious diseases in resource-poor settings. This includes recognising the biological and epidemiological aspects of tuberculosis and integrating this knowledge to inform my statistical analysis.

I have also learned to appreciate the importance of a systematic approach to data cleaning and preparation when working with secondary data. Each decision made during this phase—from handling missing values to deciding how to encode variables for analysis—had a substantial impact on the subsequent findings. This project was a testament to my ability to apply learned skills and adapt them to real-world applications, revealing both my capabilities and areas for further development. An area of self-improvement I would emphasize is the importance of self-management in the beginning phases of my research and dedicating more time to creating a comprehensive plan so that I could dedicate appropriate time to answering all aspects of the question early to avoid unforeseen issues. In particular, exploring the drug resistance variable earlier on, with its significant amount of missing data, would have allowed sufficient time to collaborate with members of the research team to link the resistance variable with an updated dataset. This would provide richer insight into the critical issue of drug-resistant tuberculosis in Peru.

Overall, the ability to formulate a question and systematically answer it using statistical and theoretical tools has been both enlightening and motivating. It has affirmed my commitment to epidemiological research and has prepared me to step into my career with confidence and a more refined skill set. This project has not only broadened my academic horizons but has also directly led to exciting career opportunities, including an invitation to join my supervisor’s lab in Peru and a scholarship by the Chadwick Trust to support my continued research. I am grateful that this experience has seamlessly connected my academic pursuits with significant growth.

# Lay Summary

Tuberculosis (TB) remains a leading cause of morbidity and mortality globally. In Lima, Peru, the scenario is particularly grave, with the second-highest incidence of TB in the Americas. This study investigates the factors influencing the radiological severity of TB, utilizing chest X-ray (CXR) as a diagnostic tool to evaluate the extent of pulmonary damage caused by the disease. This is an important area of research, as it is the first study to systematically analyze how clinical and demographic factors influence the severity of radiological pathology via chest CXR, an accessible and validated diagnostic tool in low- and middle-income countries (LMICs) like Peru. These findings can enhance the understanding of disease progression and aid in the development of targeted interventions.

This study aims to identify the clinical and demographic actors associated with worse radiological severity, measured by a validated pathology score, by carrying out an analysis of a systematic dataset with 693 participants (18 years old and above). Factors such as age, previous smoking history, and Body Mass Index (BMI) were significantly associated with CXR severity. Older age groups (58 and above) and a history of smoking were predictors of increased radiological severity, while higher BMI was inversely associated with CXR severity. Factors like socioeconomic status also became significant after looking into older age groups.

The findings of this study are pivotal in refining diagnostic and treatment strategies for TB in Peru and allow for mitigation of the disease’s impact on the most affected communities. This study contributes to the broader understanding of TB's dynamics and CXR, paving the way for enhanced disease management by the Peruvian Ministry of Health and ultimately, the reduction of TB's burden in high-prevalence settings.

# Scientific Abstract

### Background:

Tuberculosis (TB) remains the leading infectious cause of death worldwide, with Lima, Peru, experiencing a high burden of the disease. Despite the known efficacy of chest X-ray (CXR) in TB diagnosis, the impact of various factors on radiological severity is understudied. This study aimed to investigate the clinical and demographic factors associated with worse pathology scores in Lima, Peru.

### Methods:

A retrospective analysis was conducted using data from TB patients in Lima (693 patients above 18 years old), collected between 2018 and 2019. Chest x-ray (CXR) severity was quantified using a validated scoring system. A multivariate linear regression with a significance threshold of < 0.2 assessed the association between clinical and demographic variables and CXR severity scores. Sensitivity analysis was also conducted.

### Results:

The study reveals significant associations in older age groups from ages 58-67 (β = 9.3, SE = 5.25, p=0.06), ages 68-77 (β = 8.63, SE = 5.20, p=0.10), and ages 78+ (β = 13.93, SE 6.24, p=0.03). History of smoking was significantly associated (β = 4.57, SE = 3.02, p=0.13), while higher BMI was inversely associated (β = -1.56 SE = 0.29, p=8.86e-08). These findings indicate a potential long-term impact of past smoking and a potential protective effect of higher BMI on CXR pathology. Subgroup analyses within older age groups highlighted both lower socioeconomic status and injectable drug use as significant factors.

### Conclusion:

The study highlights the importance of demographic and clinical factors in the management of pulmonary TB in Peru, and the potential of CXR as a diagnostic tool in LMICs. These insights could inform targeted screening and intervention strategies from the Peruvian Ministry of Health, contributing to the global effort to curb TB by 2030 as per the United Nations Sustainable Development Goals.

# List of Acronyms

AIC - Akaike Information Criterion

BMI – Body Mass Index

CXR – Chest X-Ray

HIV – Human Immunodeficiency Virus

IQR – Interquartile Range

LMIC- Low and Middle Income Countries

MDR-TB - Multidrug-Resistant TB

NBI – Necesidades Basicas Insatisfechas (Unmet Basic Needs)

SDG - Sustainable Development Goals

SE – Standard Error

TB - Tuberculosis

UPCH - Universidad Peruana Cayetano Heredia

VIF- Variance Inflation Factor

WHO – World Health Organization

# Introduction

Tuberculosis (TB), caused by the bacteria Mycobacterium, is the most significant infectious cause of mortality and morbidity worldwide (Cormier et al., 2019). According to the World Health Organization, in 2022, an estimated 10.6 million people fell ill with TB worldwide, and 1.3 million died (WHO, 2023). Once Mycobacterium enters the lungs via inhalation of cough aerosols, it can impact several systems, such as the musculoskeletal, gastrointestinal, and central nervous; however, it is primarily a respiratory disease. The patient’s immune system responds to the bacteria by creating granulomas, a highly organized structure of many immune cells. Granulomas can either contain the infection (known as latent TB), or bacteria may disseminate to other organs and release into the respiratory tract, leading the patient to be symptomatic and infectious (known as active TB).

Pulmonary TB is hallmarked by its presentation in the lung as necrotizing granulomatous inflammation. Massive infiltration of these granulomas has been associated with increased pathology later in disease (Pichugin et al., 2009). Chest X-ray (CXR) is the primary radiological evaluation of pulmonary TB in low and middle-income settings, as it is a reliable and accessible diagnosticthat allows for the identification of pathologically advanced pulmonary TB**.** The COVID-19 pandemic has also allowed for the resurfacing of CXR techniques and emphasized the importance of CXR results in early diagnosis, treatment transition, and predicting impending complications (Gopalan et al., 2021). On CXR, TB commonly presents as a cavitary lesion in the upper lobe of the lungs in immunocompetent patients. In immunocompromised patients, such as patients with HIV, lesions can also present in the lower lobes (Mathur et al., 2017). A single, reproducible, validated numerical score for grading the extent of CXR pathology in TB to compare radiographic severity between adults with TB has also been created (Ralph et al., 2010).

Peru has the second highest incidence of TB in the Americas and has more than 40% of the multidrug-resistant TB (MDR-TB), despite making up only 3% of the population (Gianella et al., 2019). The TB Control and Prevention Act declares the fight against TB as a national interest and ending the TB epidemic by 2030 is among the health targets of the United Nations Sustainable Development Goals (SDGs). In Peru, those of poorer economic status living in the biggest cities are most affected, making it a social disease (Alarcón et al., 2017). The underutilisation of strategies that already demonstrate effectiveness, such as CXR, contributes to the slow decline of TB incidence. Active screening is one such strategy, where health systems seek out high-risk groups, and has the potential to reduce TB transmission and mortality by diagnosing and treating people earlier (Yuen et al., 2021).

In the past, the WHO positioned CXR at the end of diagnostic algorithms as a complementary tool for cases where TB is bacteriologically negative (WHO, 2016). However, there have since been advancements in radiography, which offers lower operational costs, portability, better image quality, and suitability for telemedicine, enhancing its utility in high-throughput screening and accurate TB diagnosis. Understanding predictors of worse TB pathology via CXR is a possible method to enhance active screening in Peru and LMICs, as an affordable and validated diagnostic tool.

## Scoping Literature Review

A scoping literature review was carried out to identify factors that were associated with worse TB pathology, in the context of Peru. Figure 1 below outlines the search strategy. In addition to this search engine, grey literature from the Peruvian Ministry of Health and other health organizations like the World Bank and WHO were hand-searched.

A black background with a black square

Description automatically generated with medium confidence

**Figure 1.** Search methodology for scoping literature review on factors associated with Tuberculosis pathology or disease progression.

Using these search terms, our scoping review returned 108 results, after limiting publications from 2013 to 2024. After identifying the highest impact papers, there were 78 remaining. Of the results, only 16 papers mentioned radiological TB measures, and only 9 papers were specific to TB in Peru. There are no studies that measure the association between clinical and demographic predictors and radiological pathology on chest x-ray, regardless of setting. This paper is the first study to identify the clinical and demographic factors associated with the radiological severity of TB.

A conceptual framework was developed from the existing literature, identifying key factors associated with worse TB pathology in general (see Figure 2). The factors outlined below are essential in understanding their interaction with CXR pathology, guiding our exploration of their combined impact on radiological disease severity.

A diagram of a medical procedure

Description automatically generated

**Figure 2:** Conceptual framework for factors related to worse TB pathology. (Koo et al., 2020) (du Bruyn et al., 2023) (Asgari et al., 2022) (Hargreaves et al., 2011) (Min et al., 2023) (Nhamoyebonde and Leslie, 2014).

Despite the pressing need for reliable and affordable diagnostics in Peru, a notable gap exists regarding the understanding of predictors of worse TB pathology via CXR. To address this, our study aims to identify specific clinical and demographic factors that can predict TB pathology, using a validated CXR pathology score to inform more effective strategies, such as active screening, for disease management and prevention in Peru. This is the first study to address this issue, using the only large dataset available that systematically investigates radiological TB pathology.

This study aims to understand how, and to what extent clinical and demographic factors relate to TB CXR pathology in Lima, Peru. The main objective is to identify the key factors associated with increased TB pathology CXR scores in adult men and women in the Peruvian population.

# Methods:

## Dataset and Study Population

The dataset used in this study was collected in regional labs in North and South Lima between 2018 and 2019 in patients above the age of 18. As this dataset is unpublished, access was granted via Dr. Grandjean, my supervisor. The original dataset consisted of 700 positive TB sputum samples, collected in formal clinic settings in the north and south of Lima (Lima and Callao). Dr. Sakib led the fieldwork to collect paired pre-treatment CXR, patient metadata, and CXR scoring data. All samples have been cultured, stored and linked at Universidad Peruana Cayetano Heredia (UPCH).

The CXR pathology score was calculated by two independent assessors blinded to the study results. They reviewed and scored CXRs based on severity alongside CAD4TB software, an independent, validated computer-aided detection system. The score is based on the proportion of the lung that seems affected or damaged by TB (the presence of granulomas and pulmonary cavities). Demographic and clinical variables were collected as part of the metadata collected in patient health records.

Inclusion criteria included being over the age of 18. Respondents who did not have their pulmonary score recorded were excluded. This dissertation is based on the analysis of these individuals and their samples.

## Variables for Primary Study Objective

### **Outcome Variables**

The primary outcome variable in this study is **Total Average Pathology Score**. To identify which clinical and demographic factors are more associated with worse pathology scores, a continuous variable was used on a scale from 0-100, with 100 indicating worse pathology. Participants whose pathology score was missing were removed from the original dataset and further analysis.

In sensitivity analysis, a second and separate outcome variable, **Smear Grade** was utilized. It refers to the classification of the density of TB bacteria observed in a sputum sample under a microscope, used to indicate the severity of infection and infectiousness of the patient. This is a continuous measure from 0-3, with 0 describing no bacterial load and 3 describing the highest level of bacterial load.

### **Independent Variables**

This study analyzed a wide array of independent variables to assess the association with worse TB pathology. Relevant variables were selected from the metadata provided in the original dataset and selected based on the conceptual framework. The theoretical framework behind the influence of these variables was informed based on a literature review of the existing literature (see Figure 2). Variables were categorised into two groups: clinical and demographic, which are described below.

Clinical factors include (a) **Body Mass Index (BMI)** (kg/ m²), calculated by dividing patient’s weight in kg by height in cm squared, and multiply the result by 10,000; (b) **Previous TB Diagnosis**, collected by asking patients whether they have received treatment for TB in the past (categorized as Yes or No); (c) **Diabetic**, collected by asking patients about their diabetic status (categorized as Yes, including both type 1 and type 2 combined, or No); (d) **Smoker**, collected by asking patients about their smoking status and smoking history (categorized as Yes Currently, No, or Previously); (e) **HIV Diagnosis**, collected by asking patients to reveal their HIV status (categorized as Yes or No); (f) **Injectable Drug Use**, measured by asking patients about their use of injectable drugs (categorized as Yes, No, or Previously); (g) **Sputum Grade**,which was collected by scoring sputum samples for the bacterial load under a microscope (continuous variable on a scale from 0-3, where 3 indicated worse bacterial load); and (h) **Length of Illness**, (measured in weeks; this required standardization of all responses including days, months, years, into a numerical value to represent the number of weeks).

Demographic factors included (i) **Age**, categorized into 10-year age groups (18-27, 28-37, 38-47, 48-57, 58-67, 68-77, and 78+) to translate into clinically relevant terms, age was also top coded into a 78+ category, to account for the very small number of elderly patients and to help alleviate the skewed distribution; (j) **Sex**, (categorized as Male or Female); (k) **Homelessness,** (categorised as Yes or No); (k) **Prisoner,** (categorised as Yes or No); and (l) **NBI Score,** calculated as a continuous value from 0-6, where 6 indicates worse socioeconomic position. This is calculated based on a validated method called the Necesidades Basicas Insatisfechas (NBI), a direct poverty measurement method that is constructed based on basic needs (Riofrío, 2003) (Arias and De Vos, 1996). For this method, it is necessary to define a household's or family's indispensable needs, in terms of education, health, living conditions, adequate employment, and housing services. In our dataset, we have available information on housing material, access to hygiene services, and overcrowding. Once these basic needs have been defined, a point is awarded to whether each basic need is met, and these points are summed. Individuals who have a higher score indicate a worse socioeconomic position, as they have a higher number of unmet basic needs.

## Statistical Analysis

Our primary study objective was to identify the factors associated with worse pathology scores. We divided this work into four parts: 1) collecting descriptive characteristics of the study sample to assess missingness and any potential bias as well as to determine the best model fit, 2) conducting univariate linear regression of the aforementioned variables and the outcome variable, 3) running variable multivariate linear regression model with the variables and outcome variable, and 4) conducting sensitivity analysis to assess different model structures, as well as a secondary outcome variable. Linear regressions were run based on the distribution of our outcome variable, which followed a natural distribution. Each of these steps is described in further detail below.

The first step was creating a table of the descriptive characteristics of the full study sample across all independent and dependent variables. The table includes the number and proportion of respondents to each variable (Table 1). The median was used for the continuous variables, as all variables followed some sort of skewed distribution, and is typically the best measure of central tendency for skewed data (Appendix A, Figure 1). These results also included the missingness of each variable. Complete case analysis was utilised to arrive at the analytical sample study size, excluding cases with missing variables from the dataset (see Figure 3). This was based on the low level of missingness in the data. By looking at the descriptive statistics for the analytical sample versus the full sample, little to no variation in descriptive statistics was found. A linear regression model was chosen based on the normal distribution of the outcome variable (Appendix A, Figure 2).

Second, univariate linear regression was fit to test the individual influence of each clinical and demographic variable on radiological severity, ensuring a thorough understanding of each factor prior to their collective examination in multivariate analysis. This analysis used a p-value set to 0.2. In many clinical research settings, a higher significance threshold is more favourable, considering the consequences of missing a potential predictive variable can be more severe than the risk of falsely identifying a predictive variable for worse pathology. This also allows for the inspection of the practical or clinical significance of findings, not just statistical significance.

Third, multivariate linear regression was run. The appropriateness of this model was subsequently confirmed through diagnostic checks, including Residuals vs Fitted and Normal Q-Q plots (Appendix A, Figure 3) which supported the assumptions of linearity, homoscedasticity, and normality of residuals. The VIF was also tested to identify threats of multicollinearity, and no severe threats were reported for any of our predictor variables. Therefore, the multivariate model included all predictor variables. This was also done to capture potential interactions and effects that might not be apparent in univariate analyses, thereby providing a holistic understanding of the associations. No interaction variables were included in this model either, as our study is designed to isolate and assess the strength and direction of individual factors' impacts on TB radiological severity. The multivariate model maintained a significance threshold of 0.2.

Lastly, sensitivity analysis was completed, using 4 models with a combination of variables based on the theoretical framework and the multivariate regression output to determine the best model selection and assess the robustness of findings. In all of these analyses, age and sex variables were included a priori and are included in the models regardless. This is based on the substantial literature to support the established relationship between age and sex as predicting factors in TB pathology (Humayun et al., 2022) (Min et al., 2023) (Nhamoyebonde and Leslie, 2014). These models were accompanied by AIC and adjusted R-squared values to assess model fit and explanatory power.

Analyses were performed using RStudio Server Pro Version 1.3.

## Ethics

The UPCH institutional review board granted ethical permission before the collection of the secondary dataset. Participants' consent was not obtained because the data was anonymously examined once duplicate patients were accounted for. The Peruvian Ministry of Health also granted institutional approval for the research.

# Results

## Participants

Figure 3 displays the selection criteria for the full sample after excluding respondents whose pathology score was missing (n = 1) and any outlandish values like BMI over 70 (n = 2). Patients with missing data were removed from the analysis, finalising our complete-case analytical sample (n= 563).

A black background with a black square

Description automatically generated with medium confidence

**Figure 3:** Selection criteria for analytical sample.

## Descriptive Characteristics

Males made up a higher proportion of respondents (63.91%, n = 461) and the age distribution was skewed younger, with the largest age groups being 28-37 (38.41%, n = 265), followed by 38-47 (21.3%, n = 147). The median BMI value indicates a relatively healthy weight status (22.16) with considerable variability in the BMI values (IQR = 23).

A very small proportion of patients reported the current use of injectable drugs (2.02%, n = 14), and an even smaller proportion reported using injectable drugs previously (1.73%, n = 12). A small proportion of patients reported that they were HIV positive (3.23%, n = 23). A small proportion of respondents reported being homeless, only 1.15% (n = 8).

Patients who reported smoking currently comprised 8.51% (n = 59), while previous smokers made up a larger proportion of 12.26% (n = 85). Patients who had Diabetes (type 1 or type 2) comprised 11.55% (n = 80).The median duration of illness in the sample was 4 weeks, with a broad interquartile range of nearly 6 weeks (IQR = 5.96). In total, 8.66% of patients reported receiving previous TB treatment (n = 60). The median sputum grade was 2 (IQR = 2), and the median NBI Score was 1 (IQR = 1). The proportion of missing values for most variables was small, not reaching more than 12% (NBI Score).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 1**: Descriptive characteristics for full sample | | | |  | |
| Categorical Variable | Total (n= 690)    N % | | Categorical Variable (Continued) | Total (n= 690)    N % | |
| **Age** |  |  | **prisoner** |  |  |
| 18-27 | 31 | 4.49 | Yes | 16 | 2.31 |
| 28-37 | 265 | 38.41 | No | 646 | 93.22 |
| 38-47 | 147 | 21.3 | Missing | 31 | 4.47 |
| 48-57 | 76 | 11.01 | **Diabetic** |  |  |
| 58-67 | 61 | 8.84 | Yes | 80 | 11.55 |
| 68-77 | 58 | 8.41 | No | 579 | 83.55 |
| 78+ | 28 | 4.06 | Missing | 34 | 4.91 |
| Missing | 24 | 3.48 | **Previous TB Diagnosis** |  |  |
| **Sex** |  |  | Yes | 60 | 8.66 |
| Male | 461 | 63.91 | No | 614 | 88.6 |
| Female | 231 | 32.46 | Missing | 19 | 2.74 |
| Missing | 1 | 3.62 | Continuous Variable | Estimate | IQR |
| **Injectable Drug Use** |  |  | **BMI (kg/m2)** |  |  |
| Currently | 14 | 2.02 | Median | 22.16 | 4.15 |
| Previously | 12 | 1.73 | Missing (%) | 3.48 | - |
| No | 626 | 90.33 | **Sputum grade (0-3)** |  |  |
| Missing | 41 | 5.92 | Median | 2 | 2 |
| **Homeless** |  |  | Missing (%) | 7.1 | - |
| Yes | 8 | 1.15 | **Length of illness (weeks)** |  |  |
| No | 646 | 93.22 | Median | 4 | 5.96 |
| Missing | 39 | 5.63 | Missing (%) | 3.48 | - |
| **Smoker** |  |  | **NBI Score** |  |  |
| Currently | 59 | 8.51 | Median | 1 | 4.11 |
| Previously | 85 | 12.26 | Missing (%) | 11.74 | - |
| No | 521 | 75.18 |  |  |  |
| Missing | 28 | 4.04 |  |  |  |
| **HIV diagnosis** |  |  |  |  |  |
| Yes | 23 | 3.32 |  |  |  |
| No | 634 | 91.49 | **Length of illness (weeks)** |  |  |
| Missing | 36 | 5.19 | **Diabetic** |  |  |
|  |  |  |  |  |  |

## 

## Outcome Data

In this section, the results of both univariate and multivariate linear regression analyses are presented (see Table 2), and aimed at identifying predictors of worse pathology scores for TB.

### Univariate Analysis

In univariate analysis, there was a significant increase in pathology scores in older age groups 58-67 and 78+. In ages 58-67, there was an average increase in pathology score of 7.39 points (SE=5.15, p=0.15), however, the increase was largest in patients aged 78 and above, with an average increase of 13.38 points in pathology score (SE = 6.34, p = 0.04), suggesting a significant association between older age and worse TB pathology. While no other age groups reached statistical significance, there was a trend toward higher scores in age groups above 58-67, indicating an age-related gradient in TB pathology severity.

Injectable drug use showed a significant trend toward higher pathology scores, with current users showing an estimated increase of 9.85 points (SE = 6.21, p = 0.11). However, previous users saw a rise of 6.94 points (SE = 7.06, p = 0.33), which suggests a clinically relevant but statistically insignificant relationship.

Interestingly, individuals with a history of smoking (Yes, Before) had a significant increase in pathology scores of 6.48 points (SE = 3.03, p = 0.03), indicating a potential long-term impact of past smoking on TB pathology. In contrast, current smokers had lower scores on average, although this did not reach statistical significance, with an estimate of -3.22 (SE = 3.23, p = 0.32). This counterintuitive finding could warrant further investigation into the temporal relationship between smoking cessation and TB pathology.

Continuous variables such as smear grade, length of illness, and BMI were retained in their original metric form in the univariate analysis. Smear grade was significantly associated, with a higher grade correlating to an increase of 2.06 points (SE = 0.88, p = 0.02), underscoring the importance of the clinical metric in predicting TB pathology severity. Length of illness in weeks was also significantly associated, suggesting that for each additional week of illness, the pathology score increases by 0.12 units on average (SE=0.05, p = 0.02). Furthermore, BMI was inversely related to TB pathology scores, where a higher BMI was associated with a lower score, indicated by an estimate of -1.57 points (SE = 0.27, p = 1.84e-08).

Other factors such as sex, homelessness, incarceration, diabetes, HIV status, and previous TB treatment did not show statistically significant associations in this univariate context, each with p-values well above the 0.2 threshold for significance. However, these were still included in the multivariate model to investigate their interaction with other variables further.

### Multivariate Analysis

The multivariate analysis maintained the association between older age and worse pathology scores. Individuals in the age groups above 58 years continued to exhibit a significant increase in pathology scores, with an estimated 9.83 point increase for age groups 58-67 (SE = 5.25, p = 0.06), an increase of 8.63 points in the 68-77 age group (SE = 5.20, p = 0.10), and an increase of 13.93 points in ages 78 and above (SE = 6.24, p = 0.03. These findings indicate that middle to late adulthood is an independent predictor of CXR pathology.

Injectable drug use, on the other hand, did not retain statistical significance in the multivariate analysis. However, the estimates for current and previous users remained elevated, with current users having an increase of 6.68 points (SE = 6.59, p = 0.31) and previous users demonstrating a 6.54 point increase (SE = 7.52, p = 0.38), which could be suggestive of a complex underlying relationship between drug use and TB pathology that warrants further investigation. Similarly, the length of illness in weeks did not retain statistical significance in the multivariate model, slightly surpassing the significance threshold (β = 0.07, SE = 0.05, p = 0.21).

Previous smoking presented a statistically significant estimate in the multivariate model (β = 4.57, SE = 3.02, p = 0.13), and the direction and magnitude of the association remained consistent with the univariate analysis. This reinforces the potential of an enduring impact of historical smoking on TB pathology. The inverse significant relationship between BMI and CXR pathology scores was also maintained, with an estimate of -1.56 (SE = 0.29, p = 8.66e-08), confirming nutritional status as an important factor in CXR pathology.

The multivariate analysis also showed that homelessness, with an estimate of -2.56 (SE = 7.15, p = 0.38), being a prisoner, with an estimate of -1.33 (SE = 6.63, p = 0.84), and NBI score, with an estimate of -0.18 (SE = 0.89, p = 0.84) were not significantly associated with CXR pathology scores, similar to the univariate analysis. These findings suggest that the social determinants of health may have indirect effects on TB pathology outcomes, possibly mediated through other unaddressed variables or complex social dynamics. Clinical factors, such as being diabetic (β = 1.77, SE = 3.08, p = 0.57)., having HIV (β = 1.40, SE = 5.18, p = 0.79)., and receiving previous TB treatment (β = -3.58, SE = 3.29, p = 0.28) also remained insignificant.

Graphs of the regression output (see Appendix A, Figure 3) suggest that while our multivariate regression model generally meets key assumptions, with residuals displaying random scatter and a largely normal distribution, there is some indication of potential heteroscedasticity and non-normality at the tails that warrant further investigation.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2:** Regression output for univariate and multivariate models | | | | | | | | |
|  |  | **Univariate Analysis** | | |  | **Multivariate Analysis** | | |
| **Categorical Variables** | | Estimate | Standard Error | P-value |  | Estimate | Standard Error | P-value |
| Age Group | |  |  |  |  |  |  |  |
|  | 28-37 | 3.57 | 4.36 | 0.41 |  | 3.97 | 4.28 | 0.35 |
|  | 38-47 | 4.26 | 4.58 | 0.35 |  | 4.93 | 4.56 | 0.28 |
|  | 48-57 | 1.47 | 4.95 | 0.77 |  | 4.35 | 4.93 | 0.38 |
|  | 58-67 | 7.39 | 5.15 | **0.15** |  | 9.83 | 5.25 | **0.06** |
|  | 68-77 | 6.34 | 5.15 | 0.22 |  | 8.63 | 5.20 | **0.10** |
|  | 78+ | 13.38 | 6.34 | **0.04** |  | 13.93 | 6.24 | **0.03** |
| Sex (Male) | | 0.91 | 1.96 | 0.64 |  | 0.06 | 1.97 | 0.98 |
| Injectable Drug Use | |  |  |  |  |  |  |  |
|  | Currently | 9.85 | 6.21 | **0.11** |  | 6.68 | 6.59 | 0.31 |
|  | Previously | 6.94 | 7.06 | 0.33 |  | 6.54 | 7.52 | 0.38 |
| Homeless (Yes) | | 3.57 | 9.10 | 0.70 |  | -2.56 | 9.15 | 0.78 |
| Incarcerated (Yes) | | -2.10 | 6.75 | 0.76 |  | -1.33 | 6.63 | 0.84 |
| Smoker |  |  |  |  |  |  |  |  |
|  | Currently | -3.22 | 3.23 | 0.32 |  | -4.09 | 3.36 | 0.22 |
|  | Previously | 6.48 | 3.03 | **0.03** |  | 4.57 | 3.02 | **0.13** |
| Diabetic (Yes) | | 2.68 | 2.94 | 0.36 |  | 1.77 | 3.08 | 0.57 |
| HIV (Positive) | | 1.04 | 5.05 | 0.84 |  | 1.40 | 5.18 | 0.79 |
| Previous TB Treatment (Yes) | | -3.13 | 3.28 | 0.34 |  | -3.58 | 3.29 | 0.28 |
| **Continuous Variables** | |  |  |  |  |  |  |  |
| Smear Grade | | 2.06 | 0.88 | **0.02** |  | 1.36 | 0.88 | **0.12** |
| NBI Score |  | -0.48 | 0.90 | 0.59 |  | -0.18 | 0.89 | 0.84 |
| Length of Illness (Weeks) | | 0.12 | 0.05 | **0.02** |  | 0.07 | 0.05 | 0.21 |
| BMI (kg/m²) | | -1.57 | 0.27 | **1.84e-08** |  | -1.56 | 0.29 | **8.86e-08** |
|  | | | | | | | | |

## Sensitivity Analysis

In our sensitivity analysis, four linear regression models were constructed to assess the robustness of associations with tuberculosis pathology scores (see Table 3). The first core model included age and sex, which were deemed a priori based on existing literature.

The second model expanded upon this by including variables that were statistically significant in initial univariate and multivariate analyses, such as injectable drug use (currently), smoker (currently and previously), and BMI, with the inclusion of length of illness (weeks) and smear grade, based on theoretical justification. The third model focused on a subgroup analysis for individuals aged 58 and above, with the inclusion of NBI score. Notably, the NBI score was significant in this subgroup, suggesting its potential utility in understanding pathology within an older demographic. Finally, a fourth model was explored with smear grade as the outcome variable, informed by its demonstrated correlation with TB pathology.

Each model was constructed to verify our initial findings and to explore the interactions between a different set of variables and outcomes. Adjusted R-squared values and AIC were used to evaluate model fit and complexity. They indicate that the third model has both the lowest AIC value (1119.11) and the highest adjusted R-squared (0.14). These metrics suggest the third model's potential as a robust predictor of CXR pathology when exploring older age groups.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3:** Sensitivity analysis regression output for four models | | | | | | | | | | | | | |
|  |  | **First Model (Core)** | | | **Second Model** | | | **Third Model** | | | **Fourth Model (Smear Outcome)** | | |
| **Variable** | | Estimate | Standard Error | P-value | Estimate | Standard Error | P-value | Estimate | Standard Error | P-value | Estimate | Standard Error | P-value |
| **Categorical** | |  |  |  |  |  |  |  |  |  |  |  |  |
| Age Group | |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 28-37 | 3.42 | 4.38 | 0.44 | 3.97 | 4.25 | 0.35 | - | - | - | 0.11 | 0.21 | 0.59 |
|  | 38-47 | 4.08 | 4.6 | 0.37 | 5.06 | 4.51 | 0.26 | - | - | - | 0.27 | 0.22 | 0.22 |
|  | 48-57 | 1.36 | 4.96 | 0.78 | 4.74 | 4.87 | 0.33 | - | - | - | 0.26 | 0.24 | 0.29 |
|  | 58-67 | 7.21 | 5.16 | **0.16\*** | 9.8 | 5.08 | **0.05\*** | (Reference) | - | - | 0.58 | 0.25 | **0.02\*** |
|  | 68-77 | 6.29 | 5.15 | 0.22 | 9.25 | 5.07 | **0.07\*** | -1.56 | 4.5 | 0.73 | 0.47 | 0.25 | **0.06\*** |
|  | 78+ | 13.37 | 6.34 | **0.04\*** | 14.28 | 6.19 | **0.02\*** | 5.85 | 5.81 | 0.32 | 0.45 | 0.3 | **0.14\*** |
| Sex (Male) | | 0.97 | 1.97 | 0.62 | 0.07 | 1.94 | 0.97 | 0.17 | 4.23 | 0.97 | -0.05 | 0.1 | 0.57 |
| Injectable Drug Use | |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Currently | - | - | - | 6.67 | 6.5 | 0.3 | 33.62 | 16.43 | **0.04\*** | -0.7 | 0.36 | **0.04\*** |
|  | Previously | - | - | - | 5.82 | 7.1 | 0.41 | 15.38 | 16.06 | 0.34 | -0.76 | 0.32 | **0.03\*** |
| Homeless (Yes) | | - | - | - | - | - | - | - | - | - | 1.54 | 0.44 | **5.03e−04.\*** |
| Incarcerated (Yes) | | - | - | - | - | - | - | - | - | - | -0.36 | 0.32 | 0.26 |
| Smoker |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Currently | - | - | - | -4.37 | 3.29 | **0.18\*** | 1.88 | 6.83 | 0.78 | 0.14 | 0.16 | 0.41 |
|  | Previously | - | - | - | 4.45 | 3.01 | **0.14\*** | 9.79 | 6.67 | **0.14\*** | 0.12 | 0.15 | 0.42 |
| Diabetic (Yes) | | - | - | - | - | - | - | - | - | - | 0.09 | 0.15 | 0.533 |
| HIV (Positive) | | - | - | - | - | - | - | - | - | - | 0.51 | 0.25 | **0.04\*** |
| Previous TB Treatment (Yes) | | - | - | - | - | - | - | - | - | - | -0.15 | 0.16 | 0.35 |
|  | |  |  |  |  |  |  |  |  |  |  |  |  |
| **Continuous** | |  |  |  |  |  |  |  |  |  |  |  |  |
| Smear Grade | | - | - | - | 1.42 | 0.86 | **0.1\*** | 0.73 | 1.98 | 0.71 | - | - | - |
| NBI Score |  | - | - | - | - | - | - | 3.06 | 2 | **0.13\*** | -0.02 | 0.04 | 0.62 |
| Length of Illness (Weeks) | | - | - | - | 0.06 | 0.05 | 0.25 | - | - | - | 7.20e−04 | 2.50e−03 | 0.78 |
| BMI (kg/m²) | | - | - | - | -1.57 | 0.29 | **7.13e-08\*** | -1.93 | 0.56 | **7.56e−04\*** | -0.04 | 0.02 | **8.99e−03\*** |
|  | |  | |  |  |  |  |  |  |  |  |  |  |
| Adjusted R² | | 0.00105  5095.213 | | | 0.07  5061.54 | | | 0.14  1119.11 | | | 0.04  1666.38 | | |
| AIC |  |

# Discussion

## Key Results

The primary objective of this study was to determine whether and to what extent clinical and demographic factors are associated with TB CXR pathology in adults in Lima, Peru. A major strength of this study is it is the first comprehensive risk factor analysis of radiological TB pathology. Our analysis demonstrated that middle to late adulthood is potentially critical in TB radiological pathology progression, particularly in individuals aged 58 to 78 and above. This is especially relevant for CXR as a diagnostic tool, as current evidence states the elderly often have less detectable levels of bacteria in sputum tests, contributing to delays in diagnosis when relying primarily on bacteriological testing (Li et al., 2021). Sensitivity analysis that focused on this older age group (58 – 78+) also highlighted socioeconomic status and injectable drug use as a significant risk factor, which did not exist in our primary analysis across all age groups. This highlights the importance of demographic factors in influencing TB pathology, particularly among the elderly.

This study also highlighted an inverse, significant relationship between BMI and pathology scores, where higher BMI values were associated with lower pathology scores. This underscores the potential protective effect of better nutritional status against severe TB CXR manifestations, aligning with prior findings in the literature (Humayun et al. 2022; Min et al. 2023). Results also suggest that a history of smoking, rather than current smoking status, is a significant predictor. These findings offer new insights into the temporal relationship between cigarette smoking and radiological severity, extending beyond previous studies (Quan et al., 2022),.

Sensitivity analysis using smear grade as a secondary outcome variable also confirmed significant factors like BMI and older age, validating the robustness of the associations identified in our primary analysis. It also found significant associations with a broader set of variables, including injectable drug use and HIV status. However, it did not find an association with smoking, which highlights a potential limitation of smear grade as a diagnostic and the critical gap CXR can fill in TB diagnosis and management.

## Limitations

The primary analysis found no significant effect of HIV and diabetes on radiological pathology, which could be explained by the atypical presentation of both HIV (non-cavitary disease and lower lung infiltrates) and diabetes on CXR (lower lung infiltrates) (Gopalan et al., 2021, Mathur et al., 2017, Patel et al., 2011). Furthermore, the CXR pathology scoring system used in this study, as described by Ralph et al. (2010), was initially validated in populations with a relatively low incidence of HIV co-infection, which may not fully capture the unique radiographic features associated with high HIV prevalence. Future research should aim to validate or adapt the existing CXR pathology scores in populations with higher HIV prevalence to ensure their broad applicability and accuracy in diagnosing and managing TB across different clinical scenarios.

Several other expected associations did not reach statistical significance in the multivariate regression (socioeconomic status, homelessness, incarceration, and sex), underscoring the need for further exploration of the multifactorial nature of TB pathology, perhaps considering more complex models or additional variables that might interact with the ones studied here. This could partly be explained because the measure used for socioeconomic status, NBI score, did not account for variables related to education or employment status, as they were not collected as part of the metadata. Existing literature recommends the inclusion of these variables in the construction of an NBI score (Arias and De Vos, 1996), which could mean we have an underrepresentation of the influence of an individual's socioeconomic position on their risk of TB pathology.

Furthermore, the WHO identifies alcohol use disorder as the risk factor with the most TB cases attributed in Peru, yet our study did not have a variable to measure this (WHO, 2024). This could limit our understanding of the full impact of demographic and clinical factors on TB pathology. Additionally, the study did not account for drug resistance, and given its importance in Peru, where drug resistance is a pressing concern, the exclusion might limit the applicability of our findings to populations with different drug resistance profiles. Our regression analysis also revealed potential non-linearity, outliers, and non-normal residuals, which may affect estimate precision. Despite robustness checks, these issues warrant cautious interpretation and suggest the need for advanced modelling in future research

## Interpretation

It is important to mention that this study only included adult men and women in Lima, Peru. While these results provide significant insights into TB pathology in this demographic, caution should be exercised when applying these findings to other regions or populations. Furthermore, the use of secondary data might limit the ability to capture all relevant variables that could influence TB pathology scores. However, there is great potential for comparative analysis with other LMICs with similar populations as Peru to further validate the utility of CXR for active screening, ultimately supporting a more unified global strategy against TB.

## Conclusion

* This study is the first to establish the critical role of clinical and demographic factors in influencing TB radiological pathology in Lima, Peru, highlighting specific high-risk profiles, such as older adults, patients with a history of smoking, and patients with a lower BMI, who may benefit from enhanced screening and early intervention strategies
* The findings advocate for the integration of clinical and demographic factors into TB management programs by the Peruvian Ministry of Health and highlight the importance of CXR as a diagnostic in LMICs
* Future research should aim to incorporate a wider range of demographic and clinical variables, as well as test the score in populations with higher HIV and diabetes prevalence, to fully understand the dynamics of TB pathology in different populations and enhance the effectiveness of intervention strategies

# References

ALARCÓN, V., ALARCÓN, E., FIGUEROA, C. & MENDOZA-TICONA, A. 2017. Tuberculosis in Peru: epidemiological situation, progress and challenges for its control. *Revista Peruana de Medicina Experimental y Salud Pública,* 34**,** 299-310.

ARIAS, E. & DE VOS, S. 1996. Using Housing Items to Indicate Socioeconomic Status: Latin America. *Social Indicators Research,* 38**,** 53-80.

ASGARI, S., LUO, Y., HUANG, C. C., ZHANG, Z., CALDERON, R., JIMENEZ, J., YATACO, R., CONTRERAS, C., GALEA, J. T., LECCA, L., JONES, D., MOODY, D. B., MURRAY, M. B. & RAYCHAUDHURI, S. 2022. Higher native Peruvian genetic ancestry proportion is associated with tuberculosis progression risk. *Cell Genom,* 2.

CORMIER, M., SCHWARTZMAN, K., N'DIAYE, D. S., BOONE, C. E., DOS SANTOS, A. M., GASPAR, J., CAZABON, D., GHIASI, M., KAHN, R., UPPAL, A., MORRIS, M. & OXLADE, O. 2019. Proximate determinants of tuberculosis in Indigenous peoples worldwide: a systematic review. *The Lancet Global Health,* 7**,** e68-e80.

DU BRUYN, E., STEK, C., DAROOWALA, R., SAID-HARTLEY, Q., HSIAO, M., SCHAFER, G., GOLIATH, R. T., ABRAHAMS, F., JACKSON, A., WASSERMAN, S., ALLWOOD, B. W., DAVIS, A. G., LAI, R. P. J., COUSSENS, A. K., WILKINSON, K. A., DE VRIES, J., TIFFIN, N., CERRONE, M., NTUSI, N. A. B., ABRAHAMS, F., ALLWOOD, B., AZIZ, S., BANGANI, N., BLACK, J., BLUMENTHAL, M., BREMER, M., BURGERS, W., CIKO, Z., COUSSENS, A. K., DAROOWALA, R., DU BRUYN, E., ESMAIL, H. G., GORDON, S., HARLEY, Y. X. R., HSIAO, M., LAI, R. P. J., LAKAY, F., MARTINEZ-ESTRADA, F.-O., MEINTJES, G., MENDELSON, M. S., NTUSI, N., PAPAVARNAVAS, T., PROUST, A., RUZIVE, S., SCHAFER, G., SEROLE, K., WHITAKER, C., WILKINSON, K. A., WILKINSON, R. J., ZVINAIRO, K., RIOU, C., WILKINSON, R. J. & CONSORTIUM, H. 2023. Effects of tuberculosis and/or HIV-1 infection on COVID-19 presentation and immune response in Africa. *Nature Communications,* 14**,** 188.

GIANELLA, C., PESANTES, M. A., UGARTE-GIL, C., MOORE, D. A. J. & LEMA, C. 2019. Vulnerable populations and the right to health: lessons from the Peruvian Amazon around tuberculosis control. *International Journal for Equity in Health,* 18**,** 28.

GOPALAN, N., SRINIVASALU, V. A., CHINNAYAN, P., VELAYUTHAM, B., BHASKAR, A., SANTHANAKRISHNAN, R., SENGUTTUVAN, T., RATHINAM, S., AYYAMPERUMAL, M., SATAGOPAN, K., RAJENDRAN, D., MANOHARAN, T., LAKSHMANAN, S., PARAMASIVAM, P., ANGAMUTHU, D., GANESAN, M., EASUDOSS AROCKIA, J. W., VENKATESAN, R. B., LAKSHMIPATHY, V., SHANMUGHAM, S., SUBRAMANYAM, B., SHANKAR, S., MOHIDEEN SHAHEED, J., DHANARAJ, B., PARANJI RAMIYENGAR, N., SWAMINATHAN, S. & CHANDRASEKARAN, P. 2021. Predictors of unfavorable responses to therapy in rifampicin-sensitive pulmonary tuberculosis using an integrated approach of radiological presentation and sputum mycobacterial burden. *PLoS One,* 16**,** e0257647.

HARGREAVES, J. R., BOCCIA, D., EVANS, C. A., ADATO, M., PETTICREW, M. & PORTER, J. D. 2011. The social determinants of tuberculosis: from evidence to action. *Am J Public Health,* 101**,** 654-62.

HUMAYUN, M., CHIRENDA, J., YE, W., MUKEREDZI, I., MUJURU, H. A. & YANG, Z. 2022. Effect of Gender on Clinical Presentation of Tuberculosis (TB) and Age-Specific Risk of TB, and TB-Human Immunodeficiency Virus Coinfection. *Open Forum Infect Dis,* 9**,** ofac512.

KOO, H. K., MIN, J., KIM, H. W., LEE, J., KIM, J. S., PARK, J. S. & LEE, S. S. 2020. Prediction of treatment failure and compliance in patients with tuberculosis. *BMC Infect Dis,* 20**,** 622.

LI, S.-J., LI, Y.-F., SONG, W.-M., ZHANG, Q.-Y., LIU, S.-Q., XU, T.-T., AN, Q.-Q., LIU, J.-Y. & LI, H.-C. 2021. Population aging and trends of pulmonary tuberculosis incidence in the elderly. *BMC Infectious Diseases,* 21**,** 302.

MATHUR, M., BADHAN, R. K., KUMARI, S., KAUR, N. & GUPTA, S. 2017. Radiological Manifestations of Pulmonary Tuberculosis - A Comparative Study between Immunocompromised and Immunocompetent Patients. *J Clin Diagn Res,* 11**,** Tc06-tc09.

MIN, J., PARK, J. S., KIM, H. W., KO, Y., OH, J. Y., JEONG, Y.-J., NA, J. O., KWON, S.-J., CHOE, K. H., LEE, W.-Y., LEE, S.-S., KIM, J. S. & KOO, H.-K. 2023. Differential effects of sex on tuberculosis location and severity across the lifespan. *Scientific Reports,* 13**,** 6023.

NHAMOYEBONDE, S. & LESLIE, A. 2014. Biological Differences Between the Sexes and Susceptibility to Tuberculosis. *The Journal of Infectious Diseases,* 209**,** S100-S106.

PATEL, A. K., RAMI, K. C. & GHANCHI, F. D. 2011. Radiological presentation of patients of pulmonary tuberculosis with diabetes mellitus. *Lung India,* 28**,** 70.

PICHUGIN, A. V., YAN, B. S., SLOUTSKY, A., KOBZIK, L. & KRAMNIK, I. 2009. Dominant role of the sst1 locus in pathogenesis of necrotizing lung granulomas during chronic tuberculosis infection and reactivation in genetically resistant hosts. *Am J Pathol,* 174**,** 2190-201.

QUAN, D. H., KWONG, A. J., HANSBRO, P. M. & BRITTON, W. J. 2022. No smoke without fire: the impact of cigarette smoking on the immune control of tuberculosis. *Eur Respir Rev,* 31.

RALPH, A. P., ARDIAN, M., WIGUNA, A., MAGUIRE, G. P., BECKER, N. G., DROGUMULLER, G., WILKS, M. J., WARAMORI, G., TJITRA, E., SANDJAJA, KENAGALEM, E., PONTORORING, G. J., ANSTEY, N. M. & KELLY, P. M. 2010. A simple, valid, numerical score for grading chest x-ray severity in adult smear-positive pulmonary tuberculosis. *Thorax,* 65**,** 863-9.

RIOFRÍO, G. 2003. UNDERSTANDING SLUMS: Case Studies for the Global Report on Human Settlements 2003. University College London (UCL)

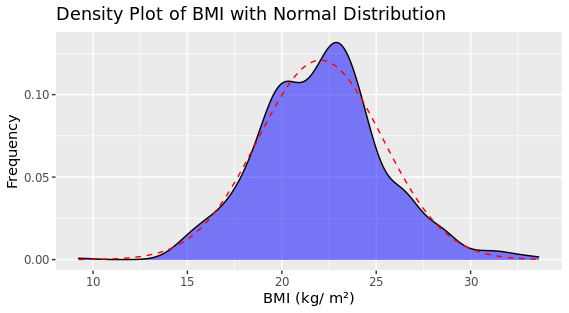
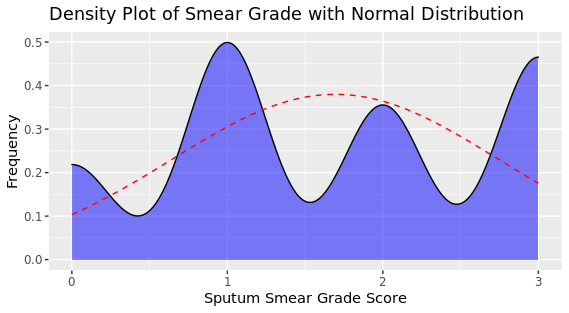
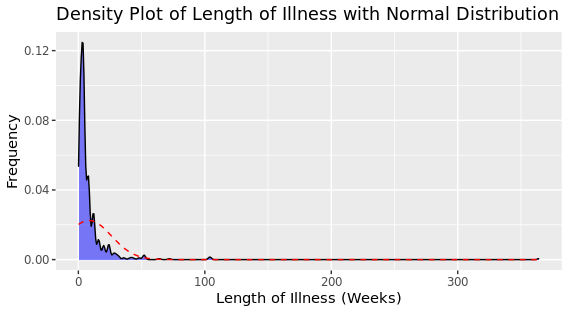
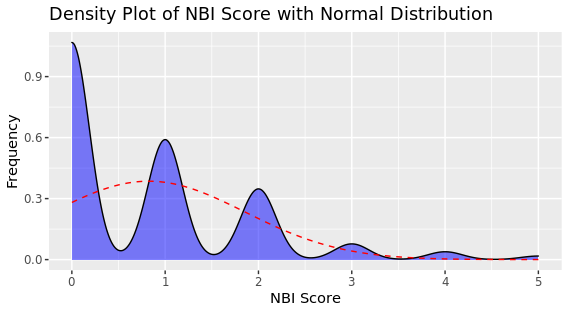
WHO 2016. Chest Radiography in Tuberculosis Detection – Summary of Current WHO Recommendations and Guidance on Programmatic Approaches.: World Health Organization

WHO. 2024. *Tuberculosis profile: Peru* [Online]. Available: https://worldhealthorg.shinyapps.io/tb\_profiles/?\_inputs\_&entity\_type=%22country%22&iso2=%22PE%22&lan=%22EN%22 [Accessed].

YUEN, C. M., PUMA, D., MILLONES, A. K., GALEA, J. T., TZELIOS, C., CALDERON, R. I., BROOKS, M. B., JIMENEZ, J., CONTRERAS, C., NICHOLS, T. C., NICHOLSON, T., LECCA, L., BECERRA, M. C. & KESHAVJEE, S. 2021. Identifying barriers and facilitators to implementation of community-based tuberculosis active case finding with mobile X-ray units in Lima, Peru: a RE-AIM evaluation. *BMJ Open,* 11**,** e050314.

# Appendix A

### Figure 1: Distribution Plots of Continuous Variables



### A graph of average distribution Description automatically generatedFigure 2: Normal Distribution Plot of Outcome Variable (Pathology Score)

### Figure 3: Multivariate Regression Output Graphs

