Sara Benist MADA Project

Tuberculosis Burden and Health Inequality Measures

Sara Benist

5/5/23

# 1. Summary/Abstract

Tuberculosis is the leading infectious cause of death, and the burden of TB is disproportionately concentrated in several high burden countries. The reason for the disparity between countries and populations of specific demographics is not fully understood. The World Health Organization collects annual data on tuberculosis outcomes through the Global Tuberculosis Programme with the goal of reducing TB incidence worldwide. To further explore the reasons behind the health inequity surrounding tuberculosis, this analysis explores the health equity indicators collected by WHO and attempts to predict tuberculosis outcomes.

The project cleans and summarizes the health equity indicators to explore disparities within the indicators based on level of income, education, sex, place of residence, and presence of drug-resistant strains. The data set was also subset into high burden countries to explore difference between high burden countries and all countries. Linear regression along with decision tree and boosted tree modelling using cross-validation tuning were conducted with predict tuberculosis outcomes and level of equity indicators.

Males have a higher proportion of cases compared to females, and level of income and education had the largest disparity between the highest and lowest levels. The prediction models did not perform well in predicting tuberculosis outcomes, most likely due to the aggregate nature of the data. More research needs to be conducted to measure the direct relationship between the health equity indicators and tuberculosis outcomes.

# 2. Introduction

## 2.1 General Background Information

Tuberculosis is an infectious disease caused by *Mycobacterium tuberculosis* and is the leading infectious cause of death in the world even though the disease is both treatable and curable(WHO, 2022c). Approximately 10 million people are infected each year, and 1.5 million die from the disease (WHO, 2022c). Humanity has been impacted by TB for centuries, but the burden of disease is not equal across countries and populations (WHO, 2022a). The World Health Organization marked 30 countries as having highest burden of TB due to the number of cases, presence of multi-drug resistance strains, and the high mortality within certain populations such as HIV (WHO, 2021c). To understand more about why some countries or groups of people experience the higher burden, the World Health Organization gathered data on several health equity indicators subgroups. The Global Tuberculosis Programme houses the End TB Strategy with a goal to eradicating TB. Their current objective is to reduce TB incidence by 80% and TB deaths by 90% by 2030 (WHO, 2022b). Previous research has focused on trends in TB outcomes; this project will explore and analyze the disparities within the indicators based on subgroup and burden of disease as well as analyzing trends in tuberculosis outcomes.

## 2.2 Description of data and data source

The World Health Organization collected data on the inequity surrounding tuberculosis, HIV, and malaria for the [State of inequality report](https://www.who.int/data/inequality-monitor/publications/report_2021_hiv_tb_malaria), and I will be exploring the dataset for TB (WHO, 2021b). More information about the data can be found [here](https://heatrepository.blob.core.windows.net/documents/data-repository-indicator-list.pdf?sp=r&st=2022-06-07T14:16:36Z&se=2023-12-30T23:16:36Z&spr=https&sv=2020-08-04&sr=b&sig=4kzThU1QDo55UOQyhWcUd8rPWJ9LxUZdRRI3zl6wKNs%3D) under “Tuberculosis Indicators” (WHO, 2021a). The dataset can be found [here](https://www.who.int/data/inequality-monitor/data#PageContent_C158_Col00) under “Tuberculosis Indicators” (WHO, 2021a).

For the tuberculosis dataset, the data was collected from the WHO Global TB programme, TB prevalence surveys, country-specific TB programmes, the WHO Health Equity Monitor database, TB patient cost surveys, and other sources (WHO, 2021b). The WHO organized the dataset to be used with the Health Equity Assessment Toolkit which is the built in data analysis and exploration tool. The database contains 10 variables regarding the burden, detection, prevention, knowledge, and social protection and observations for 194 countries over various years. The observations for each variable is further separated by up 7 inequality domains. Not all countries have data available for each year or for each inequality domain. The dataset contains a total of 7473 observations.

## 2.3 Questions/Hypotheses to be addressed

Research question: How does disparities within the health equity impact the indicator estimates? Does high disparities between inequality measures show high correlation to TB incidence, prevalence, and mortality of a country? How do these relationships differ for high burden countries compared to all countries?

The overall outcome I would study is TB mortality since effective health programs ideally reduce disease-specific mortality. The intention of this project is to identify populations that could be a focus of TB health improvement programs.

In addition to the inequality and indicator predictors provided with the dataset, I would like to examine differences in TB outcomes based on level of TB burden. This can be completed by subsetting the high burden countries into a smaller data set.

In the data, the pattern I would expect to see is a higher burden of disease on populations with greater inequality. However, I am not confident in predicting how the indicator categories, specifically the TB attitudes and perceptions, would relate to the TB outcomes.

# 3. Methods

The data cleaning will consist of widening the data set so each TB indicator is placed in a column in order to explore the indicators separately. The data set will also be split into two data sets containing low/mid burden countries (otherburden) and high burden countries (highburden). The data will then be summarized and described using exploratory analysis approaches such as graphs, plots, and tables. The statistical analysis will use decision tree and boosted tree modeling to model TB outcomes based on the available predictor variables. In order to complete the modelling, the data sets must be summarized so each country has only two values for each available indicator, one value for male and one for female. This is required because the data is aggregated, but much information on the disparities within indicators is lost.

## 3.1 Data aquisition, import, and cleaning

The WHO dataset can be found [here](o%20https://www.who.int/data/inequality-monitor/data#PageContent_C158_Col00) under “Tuberculosis Indicators”. This data was downloaded and stored into the raw\_data folder of the repository as rawdata. Information on the indicators, social determinants dimensions, and subgroups can be found in the raw\_data folder under 202206-metadata-tb.pdf. Please render the processingfile.qmd file for full results from data importing and cleaning. The associated R script file contains all coding steps with descriptions.

For the main part of the data cleaning, I used pivot\_wider() to allow each indicator to have its own column, and subset the high burden countries (as indicated by the WHO) into a new object called highburden and all other countries into otherburden. In addition, I set up the modeling data into hbmodel and obmodel which contain indicators with sex as a dimension and hbmodelprev and obmodelprev for those with place of residence. I also checked each indicator using filter(), select(), and summary() functions. These methods will allow for deeper exploration of each indicator in order to explore trends and patterns for each subgroup.

Due to the pivot\_wider() function, there was several missing observations for countries and years that had data for only some subgroups (ex: data on males but no data on females). Since I will be maintaining all the indicators in their own data set, drop\_na() would remove all data points. Instead, I will handle missing data points after exploring the data further.

The processed and cleaned data was then saved into the processed\_data folder under processeddata.rda to be used during the exploratory analysis phase. I have loaded the processed data below.

## 3.2 Statistical analysis

The exploratory analysis consisted of describing the disparities within each indicator and for the TB outcomes. This process assisted in understanding the data and framing the data for model exploration. The full exploratory analysis can be found in the exploratory\_analysis.qmd file and the associated r script.

The basic statistical analysis consisted of bivariate analyses of the TB outcomes based on subgroups identified in the exploratory analysis. For the model analysis of the data, I initially created decision tree models for the main outcomes (TB incidence, TB mortality, and TB prevalence) based on the indicator estimates for females and males and tuned the models using cross-validation folds. The data was split into test/train data sets to better evaluate model performance. This method was repeated for the otherburden and highburden countries. The full statistical analysis for decision tree modeling can be found in the statistical\_analysis.qmd file.

As a means of comparison, I also created boosted tree models for the TB outcomes following the same general methods as the decision tree modelling. The full analysis can be found in the boosted\_statistical\_analysis.qmd file.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

After processing the data, the otherburden and highburden datasets were described by exploring each health equity indicators and tuberculosis outcome. The table below shows the list of indicators, and the codebook can be found in the README.md file.

[1] "bcg" "tb\_cough" "tb\_cough\_f" "tb\_cough\_m"   
 [5] "cdr" "incidence" "mortality" "drug\_resistance"   
 [9] "tb\_att" "tb\_att\_m" "catacost" "prevalence\_place"  
[13] "p:n" "tb\_att\_f"

For each indicator, I produced a summary table using skim() to determine the number of dimensions and subgroups as well as the summary statistics to see any overall trends in the data. The summary tables are stored in the results folder.

The population for most variables were relatively low and skewed to the right. The TB incidence and mortality indicators also showed a skewed distribution. The highest incidence rate was 908 cases/100,000 , and highest mortality rate was 140 deaths/100,000.

After getting an overview of the indicators, I explored the differences between subgroups within each health equity indicator and began identifying the largest disparities. Using ggplot(), I created plots of the indicators separated by dimension and subgroup. For most variables, these disparities were described using the boxplot() function. The main outcomes of interest for TB incidence, mortality, and prevalence were plotted using the geom\_point() function.

|  |
| --- |
| Figure 1: Tuberculosis incidence by country |

[Figure 1](#fig-incexplore) shows the tuberculosis incidence for each country stratified by sex. Males have a higher proportion of TB cases, and most countries have minimal cases per 100,000 people. [Figure 2](#fig-incexplorehb) shows the TB incidence for high burden countries which are responsible for most of the TB burden worldwide. Lesotho has the highest number of cases followed by South Africa and Central African Republic.

|  |
| --- |
| Figure 2: Tuberculosis incidence by country for high burden countries |

[Figure 3](#fig-mortexplore) shows the TB mortality for high burden countries. Compared with the TB incidence values, Central African Republic has high TB incidence and mortality.

|  |
| --- |
| Figure 3: Tuberculosis mortality by country |

[Figure 4](#fig-prevexplore) shows tuberculosis prevalence by country. There may be a slight trend of urban areas having higher prevalence compared to rural areas.

|  |
| --- |
| Figure 4: Tuberculosis prevalence by country |

When exploring the health equity indicators, the BCG immunization proportion, catastrophic costs due to TB, prevalence to notification ratio, and knowledge about TB showed the largest disparities between subgroups. [Figure 5](#fig-bcg) shows the percentage of immunization for one-year-olds based on economic status, education level, area of residence, and sex. The economic status and education levels showed an approximate difference of 10-15% vaccination coverage between the highest and lowest levels.

|  |
| --- |
| Figure 5: BCG immunization by subgroup |

Experiencing catastrophic costs due to TB was most heavily present in the lowest economic quintile and was highly associated with drug resistant TB with approximately 85% of people with drug-resistant strains experiencing high costs.

|  |
| --- |
| Figure 6: Catastropic costs due to TB in highburden countries |

The prevalence to notification also shows a mild inequity between females and males as shown in [Figure 7](#fig-prevnot). Males on average have 2.4 years between contracting tuberculosis and receiving a diagnosis compared to 1.7 years for females.

|  |
| --- |
| Figure 7: Time between prevalence to notification in high burden countries |

The final indicator that showed the highest disparity is the male and female knowledge of TB shown in Figure 8. Both groups show similar disparities within the economic status, education level, and place of residence subgroups. However, females appear to have more extreme disparities within the subgroups. For example, males in the lowest quintile have 60% of the population exhibiting knowledge about how TB is spread, but females of the same quintile show only 50% of population having TB knowledge. Similar trends occur for th education and place of residence subgroups.

|  |
| --- |
| Figure 8: Knowledge about TB by subgroup for each sex |

|  |
| --- |
| Figure 9: Knowledge about TB by subgroup for each sex |

Most indicators showed the greatest disparities due to economic status and education level.

## 4.2 Basic statistical analysis

The linear models summary for TB incidence are listed below.The high burden model has the linear equation of TB incidence = 194 + 126(Male) where male would equal 1. The other burden model has the linear equation of TB incidence = 44.9 + 21.5(Male). Males having higher TB incidence is congruent with the exploratory analysis findings. The otherburden model has smaller coefficients compared to highburden model.

term estimate std.error statistic p.value  
1 (Intercept) 194.1153 24.18315 8.026883 2.492286e-12  
2 subgroupMale 126.2622 34.20014 3.691864 3.696592e-04

“Summary table for fitted linear model predicting incidence for high burden countries”

term estimate std.error statistic p.value  
1 (Intercept) 44.87916 7.318542 6.132254 2.850256e-09  
2 subgroupMale 21.58057 10.349982 2.085083 3.794231e-02

“Summary table for fitted linear model predicting incidence”

The fitted model summaries below are the TB mortality for high burden countries and other burden data. For high burden countries, the linear equation is TB mortality = 23.6 + 14.9(Male); for the other burden model, the linear equation is TB mortality = 5.36 + 3.56(Male).

term estimate std.error statistic p.value  
1 (Intercept) 23.63777 3.554384 6.650315 1.787502e-09  
2 subgroupMale 14.91024 5.026657 2.966234 3.804197e-03

“Summary table for fitted linear model predicting mortality for high burden countries”

term estimate std.error statistic p.value  
1 (Intercept) 5.362003 1.249271 4.292104 2.420371e-05  
2 subgroupMale 3.559616 1.766737 2.014797 4.485505e-02

“Summary table for fitted linear model predicting mortality”

The last outcome model I fitted was TB prevalence for high burden countries and the other burden data. For high burden countries, the linear equation is TB Prevalence = 399 + 151(Urban) with urban residences receiving a 1 and rural residences receiving a 0. The full data model has an equation of TB Prevalence = 444 - 71.4(Urban). Interestingly, the place of residence had an opposite effect on TB prevalence between the high burden and other burden models.

term estimate std.error statistic p.value  
1 (Intercept) 399.2667 53.23319 7.500333 3.606847e-08  
2 subgroupUrban 151.1333 75.28310 2.007533 5.443174e-02

“Summary table for fitted linear model predicting prevalence for high burden countries”

term estimate std.error statistic p.value  
1 (Intercept) 444.0 114.7089 3.8706677 0.00473687  
2 subgroupUrban -71.4 162.2229 -0.4401352 0.67148826

“Summary table for fitted linear model predicting prevalence”

The models can be further explored by including setting as a predictor, shown in the prediction model below as [Figure 10](#fig-prediction). The prediction model is basic prediction of TB incidence from subgroup and setting. As an example, I predicted the incidence for females and males in Central African Republic (shown in the plot). The estimates for males (602.5720; CI = 492.8893, 712.2546) and females (476.3097; CI = 366.6271, 585.9924) are shown using geom\_point() and geom\_errorbar().

|  |
| --- |
| Figure 10: Prediction plot estimating TB incidence for females and males in Central African Republic |

## 4.3 Full analysis

For the majority of the statistical analysis, the tuberculosis outcomes were modeled using the decision tree and boosted tree models and tuned with cross-validation methods from the tidymodels guide. The decision tree models all used the training data to tune the cost\_complexity and tree\_depth parameters to find the model with the lowest value of RMSE. The boosted tree models used tree\_depth, min\_n, and trees as tuning parameters.

The [Figure 11](#fig-treeinc) displays an example of the model tuning for decision tree model predicting incidence for other burden countries, and [Figure 12](#fig-treeinchb) shows the same method for high burden countries. Please refer to the statistical\_analysis.qmd for the full code and results on decision tree modelling.

|  |
| --- |
| Figure 11: Summary table for fitted tree model predicting incidence |

|  |
| --- |
| Figure 12: Summary table for fitted linear model predicting incidence for high burden countries |

A summary of the model performances are shown below predicting TB outcomes. The otherburden models are marked with ‘obdt’ and the high burden models are marked as ‘hbdt’. The outcomes being modeled are abbreviated in the model column. The best performing model predicted mortality for other burden countries (RMSE = 10.9). The TB outcome models all performed poorly in predicting TB incidence, mortality, and prevalence. The otherburden models performed better than the high burden models for incidence and mortality, and prevalence models performed worse overall.

# A tibble: 6 × 4  
 .metric .estimator .estimate model   
 <chr> <chr> <dbl> <chr>   
1 rmse standard 46.1 obdtinc   
2 rmse standard 160. hbdtinc   
3 rmse standard 9.67 obdtmort  
4 rmse standard 20.6 hbdtmort  
5 rmse standard 294. obdtprev  
6 rmse standard 190. hbdtprev

“Summary table of RMSE values for decision tree models”

The boosted tree models performed better than the decision tree models for the most part. The model section indicates the model being created. The mortality models performed best (RMSE of 8.66 and 18.5), and the prevalence models preformed significantly worse compared to the other models (RMSE of 450 and 350.7).

# A tibble: 6 × 4  
 .metric .estimator .estimate model   
 <chr> <chr> <dbl> <chr>   
1 rmse standard 39.2 obbinc   
2 rmse standard 136. hbbinc   
3 rmse standard 10.5 obbmort  
4 rmse standard 18.5 hbbmort  
5 rmse standard 450. obbprev  
6 rmse standard 351. hbbprev

“Summary table of RMSE values for boosted tree models”

# 5. Discussion

The objective of this project was to compare disparities in the health equity indicators with tuberculosis outcomes to determine if larger disparities are associated with tuberculosis burden. Due to the formatting of the data, the relationship could not be directly compared since the data is aggregated (the BCG coverage can be measured for low income individuals and females but cannot measure a low income female). Instead, each equity indicator was individually analyzed through exploratory tables and graphs, and the tuberculosis outcomes were modeled using summarised values of the indicators with sex or place of residence as dimensions.

## 5.1 Summary and Interpretation

The data for this analysis was collected by the World Health Organization to study tuberculosis outcomes and explore disparities in several underlying indicators of health equity including knowledge about how tuberculosis is spread, proportion of children vaccinated against TB, and percentage of people that experience catastrophic costs due to the disease. The health equity indicators can be viewed as potential targets of public health programs in order to reduce the morbidity of TB.

Most countries in the world experience relatively low levels of tuberculosis cases, so very few countries experience high burden of tuberculosis. In 2020, Lesotho had the highest number of TB cases (908 cases/100,000 population) followed by South Africa and Central African Republic. When considering mortality of the same year, Central African Republic had the highest mortality, indicating the country had the highest case fatality rate.

Males tended to have a higher proportion of the tuberculosis cases when exploring the incidence, prevalence, and mortality. In addition, males had a longer prevalence to notification ratio compared to females. There are many factors that may be contributing to the greater TB burden on males, this delay in notification may be associated with longer transmission time which could be a factor in the higher incidence.

The initial statistical analysis predicted tuberculosis outcomes using a linear model. The incidence and mortality models were consistent with the exploratory analysis since being male was the main predictor of the model. The high burden countries have higher levels of TB outcomes compared to the other burden countries which is consistent with the existing literature. However, these models did not take into account other variables as a predictor, so the full analysis used a decision tree model and boosted tree model to predict outcomes.

The full analysis initially consisted of a series of decision tree models that predicted TB outcomes from summarized health equity indicators. The high burden models were predicted to perform better than the other burden models, but this hypothesis did not hold true. No tuberculosis outcome models performed well, but the boosted tree models performed better compared to the decision tree models for TB incidence and mortality with high burden RMSE values of 135.5 and 18.5 and other burden RMSE values of 51 and 8.7 respectively. The prevalence model was the only outcome in which the high burden model performed better than the other burden model.

The prevalence model preformed significantly different from the incidence and mortality models. One hypothesis for the difference is the less available data for TB prevalence compared to the other outcomes or the difference in variables being used as predictors since the TB prevalence used different subgroups.

While these models are do not perform well in predicting TB outcomes, the exploratory analysis showed several populations that could be the focus of public health programs in improving health disparities. Education level and economic status showed the greatest disparity within health indicators, so public health programs can be designed to focus on improving or addressing the health equity indicators, such as hosting BCG vaccine efforts in low income populations, offering tuberculosis education courses for populations with limited educational resources, or encouraging males to be tested for TB often.

## 5.2 Strengths and Limitations

This project had several limitations associated with the analysis. Due to aggregate nature of the data and the collection years differing between indicator variables and TB outcomes, I am not able to predict tuberculosis incidence, mortality, and prevalence from the health inequity indicators directly. This aggregation is used to protect the identity of the program participants, so I do not think de-aggregated information would be publicly available. In order to complete this analysis as my research question required, I had to summarize the indicator values which lost almost all the disparities measured within the data. I believe this reduced the performance of my models because the predictors would have been more similar to each other that did not coordinate with the TB outcomes. Another limitation of the analysis is the difficulty interpreting the decision tree and boosted tree models. It was difficult to visualize and express the model due to the large number of country predictors. The final limitation is that the data did not capture all factors that contribute to the spread of tuberculosis such as population density, access to treatment regimens, consistency of adherence, and other aspects that can affect one’s susceptibility to contracting or spreading the disease. The strengths of the analysis is the substantial data set which accurately described countries of various tuberculosis burden rather than focusing specifically on countries with TB cases. The analysis also thoroughly explored each equity indicator to highlight trends and populations that could aid in the distribution of public health resources.

## 5.3 Conclusions

This analysis explores tuberculosis-related disparities between levels of income, education, sex, and other health equity topics. TB incidence and mortality is most closely associated with the country. There is no direct comparison between level of disparity in equity indicator and tuberculosis incidence, prevalence, or mortality due to the aggregated data set, so the prediction models poorly performed. Continued analysis is required to address the gap in data preventing a direct measurement of health disparity on tuberculosis outcomes. Further research should explore impact of programs addressing the health equity covered during this analysis on tuberculosis incidence and mortality.

# 6. References

WHO. (2021a). *Data*. Retrieved from <https://www.who.int/data/inequality-monitor/data#PageContent_C158_Col00>

WHO. (2021b). STATE OF INEQUALITY: HIV, TUBERCULOSIS AND MALARIA. *States News Service*. Retrieved from <https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,shib&db=edsgin&AN=edsgcl.685964785&site=eds-live&custid=uga1>

WHO. (2021c). *WHO releases new global lists of high-burden countries for TB, HIV-associated TB and drug-resistant TB*. Retrieved from <https://www.who.int/news/item/17-06-2021-who-releases-new-global-lists-of-high-burden-countries-for-tb-hiv-associated-tb-and-drug-resistant-tb>

WHO. (2022a). *Global tuberculosis report 2022*. Retrieved from <https://www.who.int/publications-detail-redirect/9789240061729>

WHO. (2022b). *The End TB Strategy*. Retrieved from <https://www.who.int/teams/global-tuberculosis-programme/the-end-tb-strategy>

WHO. (2022c). *Tuberculosis (TB)*. Retrieved from <https://www.who.int/news-room/fact-sheets/detail/tuberculosis>