Sara Benist MADA Project

Tuberculosis Burden and Health Inequality Measures

Sara Benist

2/24/23

Warning: package 'here' was built under R version 4.2.2

Warning: package 'knitr' was built under R version 4.2.2

Warning: package 'ggplot2' was built under R version 4.2.2

Warning: package 'tidyr' was built under R version 4.2.2

Warning: package 'readr' was built under R version 4.2.2

Warning: package 'purrr' was built under R version 4.2.2

Warning: package 'dplyr' was built under R version 4.2.2

Warning: package 'stringr' was built under R version 4.2.2

# 1. Summary/Abstract

Full summary to be added (TBA): background of topic, objective of project, quick overview of methods, main findings

# 2. Introduction

## 2.1 General Background Information

TBA: background of why and how I am studying this topic

## 2.2 Description of data and data source

*(Part 1 of project: data source, description, loading into raw folder, general aspects)*

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. 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”. The dataset can be found [here](o%20https://www.who.int/data/inequality-monitor/data#PageContent_C158_Col00) under “Tuberculosis Indicators”.

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. 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

*(Part 1 of project: RQ, outcomes, predictors, patterns)*

Research question: How do the TB indicators relate to the TB incidence, prevalence, and mortality of a country, and what inequality measure shows the greatest disparity in TB outcomes?

The overall outcome I would study is TB mortality since effective health programs ideally reduce disease-specific mortality. By the end of my analysis, I would like to be able to identify populations that could be a focus of TB health improvement programs. Other outcomes I would like to explore include regional differences in drug-resistant TB and the attitudes and perceptions for high burden areas compared to medium or low burden regions.

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. I would need to add another classification based on the literature to indicate what levels of TB incidence and prevalence would fall into each level of disease burden.

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

*(Part 1 of project: initial analysis thoughts)*

To analyze, I would need to determine how to handle the missing data. The years between countries are not specific, and not all indicators were collected for each country. At the moment, I would subset the data based on indicator and remove incomplete observations to analyze each indicator individually. For the final analysis, I will most likely focus on TB indicators that are significantly different between subgroups or complete enough for further analysis. As the class progresses, I look forward to learning other analysis techniques. Eventually, I would like to create a dataset that can produce the statistical information, plots, and models that explore TB outcomes based on inequality measures, TB indicators, and/or level of burden.

## 3.1 Data aquisition, import, and cleaning

*(Part 2 of project: data importing 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

#path to data  
data\_location <- here::here("data","raw\_data","202206-repository-tb.xlsx")  
  
#load data and assign to rawdata  
rawdata <- readxl::read\_excel(data\_location)  
  
str(rawdata)

tibble [7,473 × 21] (S3: tbl\_df/tbl/data.frame)  
 $ setting : chr [1:7473] "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...  
 $ year : num [1:7473] 2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...  
 $ source : chr [1:7473] "MICS" "MICS" "MICS" "MICS" ...  
 $ indicator\_abbr : chr [1:7473] "bcg" "bcg" "bcg" "bcg" ...  
 $ indicator\_name : chr [1:7473] "BCG immunization coverage among one-year-olds (%)" "BCG immunization coverage among one-year-olds (%)" "BCG immunization coverage among one-year-olds (%)" "BCG immunization coverage among one-year-olds (%)" ...  
 $ dimension : chr [1:7473] "Economic status (wealth quintile)" "Economic status (wealth quintile)" "Economic status (wealth quintile)" "Economic status (wealth quintile)" ...  
 $ subgroup : chr [1:7473] "Quintile 1 (poorest)" "Quintile 2" "Quintile 3" "Quintile 4" ...  
 $ estimate : num [1:7473] 53.8 61.9 58.1 65.1 77.9 ...  
 $ se : num [1:7473] 4.2 3.12 3.37 3.77 2.24 ...  
 $ ci\_lb : num [1:7473] 45.5 55.6 51.4 57.4 73.1 ...  
 $ ci\_ub : num [1:7473] 61.8 67.8 64.6 72.2 81.9 ...  
 $ population : num [1:7473] 532 549 495 473 447 ...  
 $ flag : chr [1:7473] NA NA NA NA ...  
 $ setting\_average : num [1:7473] 62.9 62.9 62.9 62.9 62.9 ...  
 $ iso3 : chr [1:7473] "AFG" "AFG" "AFG" "AFG" ...  
 $ favourable\_indicator: num [1:7473] 1 1 1 1 1 1 1 1 1 1 ...  
 $ indicator\_scale : num [1:7473] 100 100 100 100 100 100 100 100 100 100 ...  
 $ ordered\_dimension : num [1:7473] 1 1 1 1 1 1 1 1 0 0 ...  
 $ subgroup\_order : num [1:7473] 1 2 3 4 5 1 2 3 0 0 ...  
 $ reference\_subgroup : num [1:7473] 0 0 0 0 0 0 0 0 0 1 ...  
 $ topic : chr [1:7473] "TB" "TB" "TB" "TB" ...

Please see processingfile.qmd for full code describing data importing and cleaning.

The rawdata file contains information on the country, year of data collection, the indicator they are studying, the social determinants of health they are considering, and the subgroup of each dimension. The dataset is mostly complete with only a few main variables missing data points

#another view of the data  
skimr::skim(rawdata)

Data summary

|  |  |
| --- | --- |
| Name | rawdata |
| Number of rows | 7473 |
| Number of columns | 21 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 9 |
| numeric | 12 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| setting | 0 | 1.00 | 4 | 37 | 0 | 194 | 0 |
| source | 0 | 1.00 | 3 | 61 | 0 | 8 | 0 |
| indicator\_abbr | 0 | 1.00 | 3 | 16 | 0 | 14 | 0 |
| indicator\_name | 0 | 1.00 | 23 | 67 | 0 | 14 | 0 |
| dimension | 0 | 1.00 | 3 | 33 | 0 | 7 | 0 |
| subgroup | 0 | 1.00 | 4 | 29 | 0 | 19 | 0 |
| flag | 7113 | 0.05 | 30 | 55 | 0 | 3 | 0 |
| iso3 | 0 | 1.00 | 3 | 3 | 0 | 194 | 0 |
| topic | 0 | 1.00 | 2 | 2 | 0 | 1 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| year | 0 | 1.00 | 2010.88 | 7.40 | 1991 | 2006.00 | 2012.00 | 2017.00 | 2020.00 | ▁▂▅▆▇ |
| estimate | 132 | 0.98 | 74.18 | 56.58 | 0 | 48.01 | 84.09 | 95.34 | 1014.00 | ▇▁▁▁▁ |
| se | 1705 | 0.77 | 1.71 | 1.21 | 0 | 0.92 | 1.46 | 2.20 | 14.07 | ▇▁▁▁▁ |
| ci\_lb | 824 | 0.89 | 66.89 | 40.77 | 0 | 43.26 | 79.97 | 91.51 | 738.00 | ▇▁▁▁▁ |
| ci\_ub | 824 | 0.89 | 90.07 | 106.22 | 0 | 64.56 | 91.78 | 98.18 | 2176.07 | ▇▁▁▁▁ |
| population | 287 | 0.96 | 2164009.02 | 25032321.78 | 18 | 300.80 | 752.27 | 3212.95 | 738247340.00 | ▇▁▁▁▁ |
| setting\_average | 294 | 0.96 | 76.35 | 53.65 | 0 | 56.01 | 84.75 | 94.98 | 831.00 | ▇▁▁▁▁ |
| favourable\_indicator | 0 | 1.00 | 0.76 | 0.43 | 0 | 1.00 | 1.00 | 1.00 | 1.00 | ▂▁▁▁▇ |
| indicator\_scale | 0 | 1.00 | 10525.91 | 30467.66 | 10 | 100.00 | 100.00 | 100.00 | 100000.00 | ▇▁▁▁▁ |
| ordered\_dimension | 0 | 1.00 | 0.55 | 0.50 | 0 | 0.00 | 1.00 | 1.00 | 1.00 | ▆▁▁▁▇ |
| subgroup\_order | 0 | 1.00 | 1.42 | 1.60 | 0 | 0.00 | 1.00 | 3.00 | 5.00 | ▇▂▂▁▁ |
| reference\_subgroup | 0 | 1.00 | 0.22 | 0.41 | 0 | 0.00 | 0.00 | 0.00 | 1.00 | ▇▁▁▁▂ |

The indicators being studied are shown below.

#look at what indicators are in the dataset  
unique(rawdata$indicator\_name)

[1] "BCG immunization coverage among one-year-olds (%)"   
 [2] "People who report TB is spread through coughing (%)"   
 [3] "People who report TB is spread through coughing - Female (%)"   
 [4] "People who report TB is spread through coughing - Male (%)"   
 [5] "Case detection rate (%)"   
 [6] "TB incidence (new infections per 100 000 population)"   
 [7] "TB mortality (deaths per 100 000 population)"   
 [8] "People with MDR/RR-TB (%)"   
 [9] "People who would want a family member's TB kept secret (%)"   
[10] "People who would want a family member's TB kept secret - Male (%)"   
[11] "TB prevalence (cases per 100 000 population)"   
[12] "Prevalence to notification ratio (years)"   
[13] "Families affected by TB facing catastrophic costs due to TB (%)"   
[14] "People who would want a family member's TB kept secret - Female (%)"

For the main part of the data cleaning, I used pivot\_wider() to allow each indicator to have its own column, and assign each indicator into its own object using filter() and select() functions. These methods will allow for deeper exploration of each indicator in order to explore trends and patterns for each subgroup. An example of the subsetting is shown below.

wide\_data <- rawdata %>%   
 select(c(setting, year, indicator\_name,  
 indicator\_abbr, dimension, subgroup,  
 estimate, population)) %>%   
 pivot\_wider(names\_from = "indicator\_name", values\_from = "estimate")  
slice(wide\_data)

# A tibble: 7,473 × 20  
 setting year indic…¹ dimen…² subgr…³ popul…⁴ BCG i…⁵ Peopl…⁶ Peopl…⁷ Peopl…⁸  
 <chr> <dbl> <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
 1 Afghan… 2010 bcg Econom… Quinti… 532. 53.8 NA NA NA  
 2 Afghan… 2010 bcg Econom… Quinti… 549. 61.9 NA NA NA  
 3 Afghan… 2010 bcg Econom… Quinti… 495. 58.1 NA NA NA  
 4 Afghan… 2010 bcg Econom… Quinti… 473. 65.1 NA NA NA  
 5 Afghan… 2010 bcg Econom… Quinti… 447. 77.9 NA NA NA  
 6 Afghan… 2010 bcg Educat… No edu… 2267. 61.1 NA NA NA  
 7 Afghan… 2010 bcg Educat… Primar… 122. 76.3 NA NA NA  
 8 Afghan… 2010 bcg Educat… Second… 108. 85.9 NA NA NA  
 9 Afghan… 2010 bcg Place … Rural 2060. 59.8 NA NA NA  
10 Afghan… 2010 bcg Place … Urban 436. 77.3 NA NA NA  
# … with 7,463 more rows, 10 more variables: `Case detection rate (%)` <dbl>,  
# `TB incidence (new infections per 100 000 population)` <dbl>,  
# `TB mortality (deaths per 100 000 population)` <dbl>,  
# `People with MDR/RR-TB (%)` <dbl>,  
# `People who would want a family member's TB kept secret (%)` <dbl>,  
# `People who would want a family member's TB kept secret - Male (%)` <dbl>,  
# `TB prevalence (cases per 100 000 population)` <dbl>, …

# BCG coverage indicator  
bcg <- wide\_data %>%   
 filter(indicator\_abbr == "bcg") %>%   
 select(c(1,2,3,4,5,6,7))  
summary(bcg)

setting year indicator\_abbr dimension   
 Length:4352 Min. :1991 Length:4352 Length:4352   
 Class :character 1st Qu.:2003 Class :character Class :character   
 Mode :character Median :2009 Mode :character Mode :character   
 Mean :2008   
 3rd Qu.:2014   
 Max. :2019   
   
 subgroup population   
 Length:4352 Min. : 18.9   
 Class :character 1st Qu.: 210.3   
 Mode :character Median : 377.8   
 Mean : 1666.4   
 3rd Qu.: 719.8   
 Max. :696209.9   
 NA's :96   
 BCG immunization coverage among one-year-olds (%)  
 Min. : 5.57   
 1st Qu.: 85.97   
 Median : 93.65   
 Mean : 88.73   
 3rd Qu.: 97.25   
 Max. :100.00   
 NA's :96

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). Most of these observations were dropped since no important information was lost by removing these “empty” observations. An example is shown below.

# BCG missing population & bcg data   
# due to missing info for dimension levels  
bcg <- bcg %>%   
 drop\_na()

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

## 3.2 Statistical analysis

TBA: statistical analysis description

# 4. Results

## 4.1 Exploratory/Descriptive analysis

*(Part 2 of project: exploratory analysis)* Please see exploratory\_analysis.qmd for full code describing data exploratory analysis.

The processedcode was loaded into the exploratory\_analysis.qmd file which pulls coding from the exploratoryanalysis.r script.

#Path to data.  
data\_location <- here::here("data","processed\_data","processeddata.rda")  
#load data  
load(data\_location)

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. An example of the code is shown below.

#incidence summary and save to file location  
summary\_inc <- skimr::skim(incidence)  
print(summary\_inc)

── Data Summary ────────────────────────  
 Values   
Name incidence  
Number of rows 388   
Number of columns 7   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Column type frequency:   
 character 4   
 numeric 3   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Group variables None   
  
── Variable type: character ────────────────────────────────────────────────────  
 skim\_variable n\_missing complete\_rate min max empty n\_unique whitespace  
1 setting 0 1 4 37 0 194 0  
2 indicator\_abbr 0 1 9 9 0 1 0  
3 dimension 0 1 3 3 0 1 0  
4 subgroup 0 1 4 6 0 2 0  
  
── Variable type: numeric ──────────────────────────────────────────────────────  
 skim\_variable n\_missing complete\_rate  
1 year 0 1  
2 population 0 1  
3 TB incidence (new infections per 100 000 population) 0 1  
 mean sd p0 p25 p50 p75 p100 hist   
1 2020 0 2020 2020 2020 2020 2020 ▁▁▇▁▁  
2 19974255. 73893692. 798 1013975. 4369962. 14508898 738247340 ▇▁▁▁▁  
3 107. 147. 0 9.28 40.9 146. 908. ▇▁▁▁▁

incsummarytable\_file <- here("results", "summary tables", "incsummarytable.rds")  
saveRDS(summary\_inc, file = incsummarytable\_file)  
  
#mortality summary and save to file location  
summary\_mort <- skimr::skim(mortality)  
print(summary\_mort)

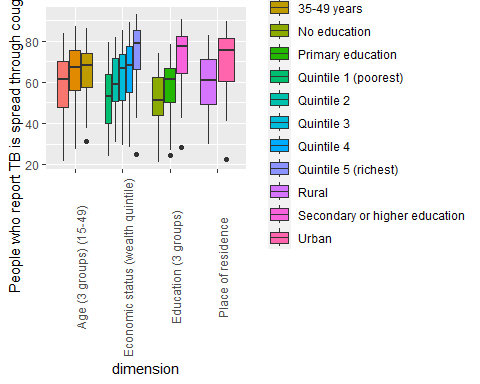
── Data Summary ────────────────────────  
 Values   
Name mortality  
Number of rows 388   
Number of columns 7   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Column type frequency:   
 character 4   
 numeric 3   
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_   
Group variables None   
  
── Variable type: character ────────────────────────────────────────────────────  
 skim\_variable n\_missing complete\_rate min max empty n\_unique whitespace  
1 setting 0 1 4 37 0 194 0  
2 indicator\_abbr 0 1 9 9 0 1 0  
3 dimension 0 1 3 3 0 1 0  
4 subgroup 0 1 4 6 0 2 0  
  
── Variable type: numeric ──────────────────────────────────────────────────────  
 skim\_variable n\_missing complete\_rate  
1 year 0 1  
2 population 0 1  
3 TB mortality (deaths per 100 000 population) 0 1  
 mean sd p0 p25 p50 p75 p100 hist   
1 2020 0 2020 2020 2020 2020 2020 ▁▁▇▁▁  
2 19974255. 73893692. 798 1013975. 4369962. 14508898 738247340 ▇▁▁▁▁  
3 13.2 21.1 0 0.630 3.39 16.9 140. ▇▁▁▁▁

mortsummarytable\_file <- here("results", "summary tables", "mortsummarytable.rds")  
saveRDS(summary\_mort, file = mortsummarytable\_file)

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.

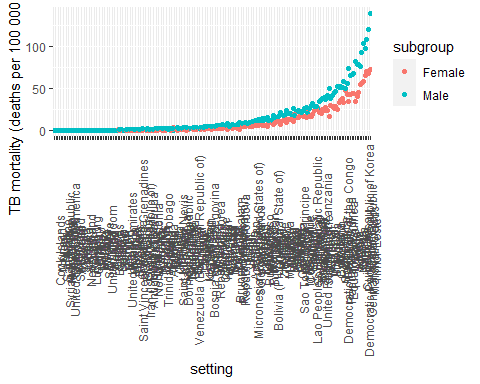
After getting an overview of the indicators, I wanted to explore the differences between subgroups and start identifying the largest disparities. Using ggplot(), I created plots of the indicators separated by dimension and subgroup. For most variables, these disparities were described using the boxplot() function as shown below.

p13 <- tb\_cough\_f %>% ggplot(aes(x=dimension,  
 y = `People who report TB is spread through coughing - Female (%)`,  
 fill = subgroup)) +  
 geom\_boxplot()+  
 theme(axis.text.x = element\_text(angle = 90))   
plot(p13) # plots knowledge about TB by dimension and subgroup for females



The main outcomes of interest for TB incidence, mortality, and prevalence were plotted using the geom\_point() function as shown below.

p6 <- mortality %>% ggplot(aes(x=fct\_reorder(  
 setting, `TB mortality (deaths per 100 000 population)`),  
 y = `TB mortality (deaths per 100 000 population)`,  
 color = subgroup)) +  
 geom\_point()+  
 theme(axis.text.x = element\_text(angle = 90))+  
 scale\_fill\_brewer(palette = "Spectral")+  
 xlab("setting")  
plot(p6) #plots mortality by country from lowest mortality to highest, colored by subgroup



Most indicators showed the greatest disparities due to economic status and education level. Further exploratory analysis should look into level of TB burden as a possible variable that should be considered, and statistical analysis in the next step should attempt to quantify the disparities between subgroups and the impact on TB outcomes.

## 4.2 Basic statistical analysis

TBA: simple statistical models for associations in data

## 4.3 Full analysis

TBA: final analysis with statistic/ML methods, meaningful figures/tables, load saved products here

# 5. Discussion

## 5.1 Summary and Interpretation

TBA: summary of findings and why it matters

## 5.2 Strengths and Limitations

TBA: strengths and limitations of analysis

## 5.3 Conclusions

TBA: take away messages, citations from bibtex

# 6. References