

**TB Burden Analysis by Country**

*An Interactive Data Visualization Application Using R Shiny*

**Niranjan Rao**

San Jose State University

DATA 230 - Business Intelligence and Data Visualization

**Date:** November 12, 2024

## Title: TB Burden Analysis by Country – R Shiny Application

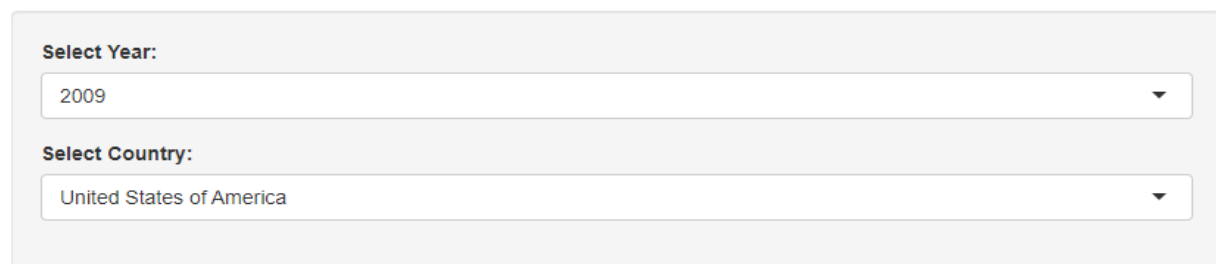
### Introduction

This report presents an interactive R Shiny application designed to visualize and analyze tuberculosis (TB) burden data by country. The application includes three main visualizations:

1. **Top 20 Countries by TB Prevalence Bar Plot**
2. **TB Prevalence Over Time Line Plot** for a selected country
3. **TB Prevalence vs. Case Detection Rate Scatter Plot** for a selected year

Each visualization serves a unique purpose, allowing users to explore TB prevalence trends, compare countries, and understand the relationship between TB detection and prevalence.

## TB Burden Analysis by Country



The screenshot displays the user interface of the R Shiny application. It features two dropdown menus within a light gray container. The first menu, labeled 'Select Year:', has '2009' selected. The second menu, labeled 'Select Country:', has 'United States of America' selected. Both menus have a small downward arrow on the right side of the selection box.

### 1. Top 20 Countries by TB Prevalence (Bar Plot)

The bar plot displays the top 20 countries with the highest TB prevalence for the selected year. The code performs the following steps:

- **Data Filtering:** The data is filtered based on the selected year to focus on that specific snapshot.
- **Aggregation:** The code groups the data by Country.or.territory.name and calculates the average TB prevalence per 100,000 people.
- **Sorting:** The top 20 countries with the highest TB prevalence are selected by sorting in descending order.
- **Plot Customization:** A bar plot is created with a gradient fill from light to dark blue, enhancing the visual impact. The bars are arranged horizontally for readability, with the highest prevalence at the top.

### Insights:

This plot helps identify countries with the most significant TB burdens in a particular year. The color gradient visually emphasizes countries with extremely high prevalence, allowing for quick comparisons.

```
ggplot(top_countries, aes(x = reorder(Country.or.territory.name, TB_Prevalence), y =
TB_Prevalence, fill = TB_Prevalence)) +

  geom_bar(stat = "identity") +

  coord_flip() + # Flip for easier reading

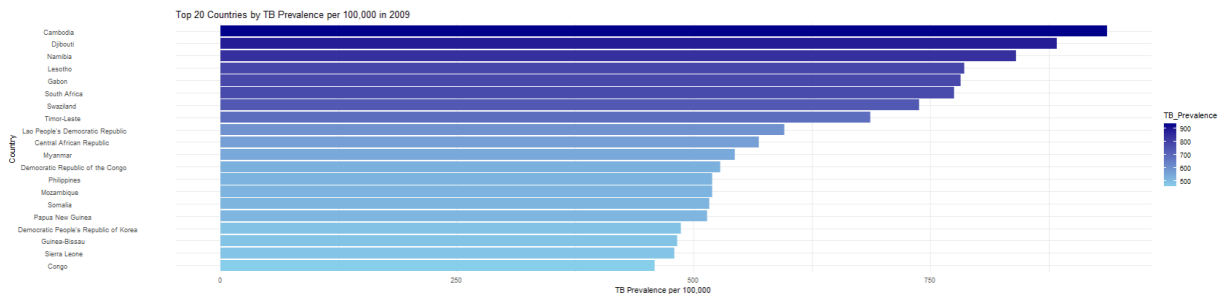
  scale_fill_gradient(low = "skyblue", high = "darkblue") +

  ggtitle(paste("Top 20 Countries by TB Prevalence per 100,000 in", input$year)) +

  xlab("Country") +

  ylab("TB Prevalence per 100,000") +

  theme_minimal()
```



## 2. TB Prevalence Over Time (Line Plot)

The line plot shows TB prevalence over time for a user-selected country, helping visualize trends in a single location.

- **Data Filtering:** The data is filtered for the chosen country across all available years.
- **Aggregation:** TB prevalence is averaged for each year to ensure consistency.
- **Plot Customization:** A line plot is created with a blue line and point markers, visually tracking TB prevalence changes. This color choice makes the trend easy to follow, and the markers highlight specific data points.

### Insights:

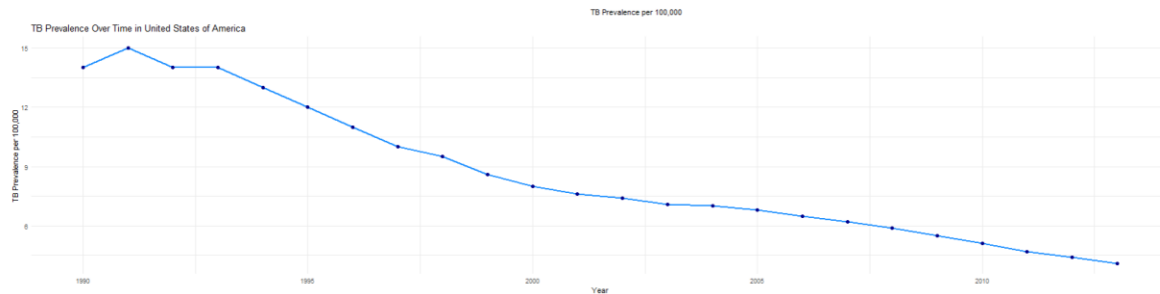
This plot is valuable for understanding trends within a country, indicating whether TB prevalence is increasing, decreasing, or remaining stable. For example, a steady decline might suggest effective public health interventions, whereas an upward trend may signal emerging issues in TB management.

```

country_data <- data %>%
  filter(Country.or.territory.name == input$country) %>%
  group_by(Year) %>%
  summarize(TB_Prevalence =
    mean(Estimated.prevalence.of.TB..all.forms..per.100.000.population, na.rm = TRUE))

ggplot(country_data, aes(x = Year, y = TB_Prevalence)) +
  geom_line(color = "dodgerblue", size = 1) +
  geom_point(color = "darkblue", size = 2) +
  ggtitle(paste("TB Prevalence Over Time in", input$country)) +
  xlab("Year") +
  ylab("TB Prevalence per 100,000") +
  theme_minimal()

```



### 3. TB Prevalence vs. Case Detection Rate (Scatter Plot)

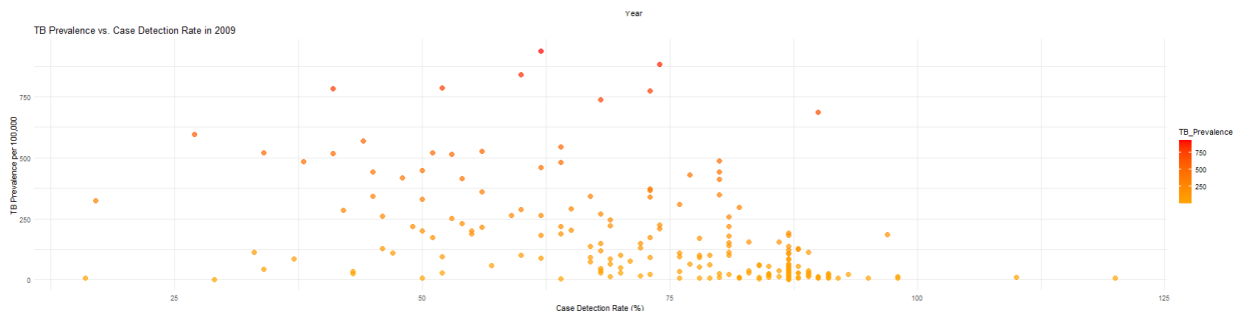
The scatter plot compares TB prevalence to the case detection rate for each country in a selected year, providing insights into the relationship between detection rates and TB burden.

- **Data Filtering:** The data is filtered to include only the selected year.
- **Variable Selection:** The scatter plot uses TB prevalence and case detection rate as the x and y variables, respectively.
- **Plot Customization:** Each point represents a country and is colored based on TB prevalence, with a gradient from orange to red to indicate lower to higher values. The color gradient enhances visual separation of high-prevalence countries.

## Insights:

This plot helps assess whether countries with higher case detection rates have lower or higher TB prevalence. For example, if a country with a low case detection rate has a high TB prevalence, it may indicate under-detection of cases, requiring improved screening and diagnostic efforts.

```
scatter_data <- data %>%  
  filter(Year == input$year) %>%  
  select(Country.or.territory.name, TB_Prevalence =  
Estimated.prevalence.of.TB..all.forms..per.100.000.population,  
         Case_Detection = Case.detection.rate..all.forms...percent) %>%  
  na.omit() # Remove rows with NA values  
  
ggplot(scatter_data, aes(x = Case_Detection, y = TB_Prevalence, color = TB_Prevalence)) +  
  geom_point(size = 3, alpha = 0.7) +  
  scale_color_gradient(low = "orange", high = "red") +  
  ggtitle(paste("TB Prevalence vs. Case Detection Rate in", input$year)) +  
  xlab("Case Detection Rate (%)") +  
  ylab("TB Prevalence per 100,000") +  
  theme_minimal()
```



## Application Interface and User Interaction

The Shiny application interface consists of a sidebar and a main panel:

- **Sidebar Inputs:**

- **Year Selector:** A dropdown menu for selecting the year, affecting the bar and scatter plots.
- **Country Selector:** Allows users to select a specific country, affecting the line plot showing TB prevalence over time.
- **Main Panel:** Displays the three interactive plots, updating dynamically based on user input.

### Code Workflow in the Shiny Application

- **Reactive Expressions:** The application uses reactive expressions to filter and aggregate data based on user-selected inputs. Each plot is rendered in response to changes in the selected year or country.
- **Render Functions:** Each plot has a corresponding renderPlot function, which contains the logic for data filtering, aggregation, and plot customization.

### Conclusion

This Shiny application provides an intuitive way to explore TB burden data across countries and years. The interactive plots enable users to:

1. Identify high-prevalence countries in a given year.
2. Track TB trends within a country over time.
3. Analyze the relationship between TB detection efforts and TB prevalence.

With these visualizations, public health professionals and researchers can gain valuable insights into TB distribution and the effectiveness of case detection, guiding efforts to manage and reduce TB prevalence globally.

## References

1. RStudio, PBC. (2021). *Shiny: Web Application Framework for R*. Retrieved from <https://shiny.rstudio.com>
2. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
3. Natural Earth. (2021). *Free vector and raster map data*. Retrieved from <https://www.naturalearthdata.com>
4. World Health Organization. (2023). *Global tuberculosis report 2023*. Retrieved from <https://www.who.int/teams/global-tuberculosis-programme>
5. Grolemund, G., & Wickham, H. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.