MTH208: Course Project Report

Group 24

Global Tuberculosis Analysis

Under the supervision of: Professor Akash Anand

Team Members:

Naina Bhalla (240674) Ananya Aggarwal (251080056)

Puspak Kumar Laha (251080084) Ravi Kumar (251080085)

Contents

Acknowledgement

We would like to express our sincere gratitude to Prof. Akash Anand for his valuable guidance and support throughout this project. Working on this topic provided us with an opportunity to apply data science techniques to an issue of global importance and deepened our understanding of data analysis, visualization, and interpretation. We are also extremely grateful for the constant cooperation and dedication of all team members, whose efforts made this project a successful and enriching experience.

1 Introduction

Tuberculosis (TB) is an infectious disease caused by Mycobacterium tuberculosis, primarily affecting the lungs, and remains one of the leading causes of death from a single infectious agent worldwide. India bears a disproportionately high burden of TB, with mortality rates remaining alarmingly high despite national and global efforts to control the disease. The prevalence and severity of TB are closely linked to social and economic factors, including poverty, under nutrition, literacy, urbanization, and access to healthcare.

Understanding these patterns is essential for addressing the disease effectively. By examining TB trends and their relation to socioeconomic factors, we can gain valuable insights into its persistence and the populations most at risk, which is critical for guiding public health strategies and interventions.

2 Objectives

- To analyze global and country-specific trends in TB incidence, mortality, and notification rates over time, identifying variations across regions and income groups.
- To investigate the association between TB burden and socioeconomic determinants in order to understand how different influence disease dynamics.
- To examine patterns and trends of drug-resistant tuberculosis, including multidrug-resistant (MDR-TB) and rifampicin-resistant (RR-TB) cases
- To assess how the disruptions in healthcare services during the Covid-19 pandemic affected Global TB notifications and incidence reporting.
- To design and implement a user-friendly Shiny dashboard that allows interactive exploration of global TB data
- To support evidence-based decision-making by providing a research-oriented tool that can be extended for policy evaluation, academic research, and public health planning.

3 Data Collection

3.1 Sources

- **WHO TB Database:**
  Extracted directly fromWHO Global TB Programme Datasets (scraped from the WHO data portal, 2025 release).
- **World Development Indicators (WDI):**
  Downloaded via the `WDI` R package, which accesses the World Bank's open API.

3.2 Data Description

The dataset integrates epidemiological, financial, drug-resistance, socioeconomic, health and environmental indicators relevant to tuberculosis (TB). Variables follow the naming and selection implemented in the project scripts; key domains and columns are:

1. **Epidemiological indicators (WHO burden files):**
   - `country`, `year`, `g_whoregion`: WHO region classification.
   - `e_pop_num` — estimated total population.
   - `e_inc_100k`, `e_inc_num`: estimated TB incidence (per 100k and absolute).
   - `e_mort_100k`, `e_mort_num`: estimated TB mortality (per 100k and absolute).
   - `e_inc_tbhiv_100k`, `e_inc_tbhiv_num`: estimated TB incidence among people living with HIV.
   - `e_mort_tbhiv_100k`, `e_mort_tbhiv_num`: estimated TB mortality among people living with HIV.
   - `cfr`, `cfr_pct`: case fatality ratio (raw and %).
   - `c_newinc_100k`: notifications (new + relapse) per 100k.
   - `c_cdr`: case detection rate (proportion of estimated cases that are reported).
2. **Financial / budget indicators (WHO budget files):**
   - `budget_tot`: total budget (where available).
   - `cf_tot_sources`: total confirmed funding sources.
   - `budget_lab`: laboratory budget.
   - `budget_tpt`: budget for preventive treatment.
   - `budget_mdrmgt` (or `budget_mdrmgt`): budget for MDR-TB management.
   - `cf_tot_domestic`, `cf_tot_gf`, `cf_tot_usaid` : domestic, Global Fund, USAID contributions.
3. **Drug-resistant TB indicators (MDR/RR burden files):**
   - `e_rr_pct_new`: % rifampicin-resistant among new cases.
   - `e_rr_pct_ret`: % rifampicin-resistant among retreatment cases.
   - `e_inc_rr_num`: estimated number of RR-TB cases.

4. **Socioeconomic indicators (World Bank WDI codes pulled via `WDI`):**
    - `NY.GDP.PCAP.CD`: GDP per capita (current US$).
    - `SP.POP.TOTL`: total population.
    - `SP.URB.TOTL.IN.ZS`: urban population (%).
    - `SH.XPD.CHEX.PC.CD`: health expenditure per capita (current US$).
    - `SI.POV.DDAY`: poverty headcount ratio.
    - `SE.ADT.LITR.ZS`: adult literacy rate (%).
5. **Health-related risk factors (WDI):**
    - `SH.DYN.AIDS.ZS`: HIV prevalence (%).
    - `SH.STA.DIAB.ZS`: diabetes prevalence (%).
    - `SH.PRV.SMOK`: smoking prevalence (%).
6. **Environmental indicators (WDI):**
    - `EN.ATM.PM25.MC.M3`: PM2.5 mean exposure (μg/m^3).
    - `EN.ATM.PM25.MC.ZS`: % population above WHO PM2.5 guideline.
7. **Country identification:**
    - `iso2c`, `iso3c`: ISO2 / ISO3 country codes (added with `countrycode`).
    - `country_who`, `country_wdi`:country names as found in WHO and WDI sources after merging/renaming.

3.3 Preprocessing

The preprocessing implemented in the project scripts proceeds as follows (script references given for reproducibility):

1. **Read WHO source files.** The scraping scripts read CSV snapshots extracted from WHO reports (files under `data/who_data/`). The code selects only the columns required for analysis (see `scraping.‘`)
2. **Select and join budget and MDR/RR burden tables.** Budget fields are left-joined to the main WHO burden table by `country` and `year` (implemented with `dplyr::left_join` in `scraping.R`).
3. **Download and attach World Bank (WDI) indicators.**The script loads a hand-picked set of WDI indicators using

```
WDI::WDI(..., start = 2000, end =
    2023)
```

for socioeconomic and environmental indicators. They are combined into a single WDI table: (`wdi_full`).

4. **Add ISO codes and merge WHO + WDI.** Country ISO2/ISO3 codes are created with

```
countrycode::countrycode()
```

and appended to the WHO table. The WHO table is then left-joined with the WDI table by `iso3c` and `year`. The combined object is renamed (e.g., `country_who` / `country_wdi`) and filtered to remove rows missing key identifiers.

5. **Missing-value handling.** Missing values are generally handled by filtering or by aggregation functions with `na.rm=TRUE` in the plotting scripts (`home_plots.R`, `global_plots.R`).
6. **Save final dataset.** The final merged dataset is written to `data/who_tb_global.csv` (this is the relative path used by the scraping scripts). When running the scraping code from `app/` the file is written under `app/data/who_tb_global.csv` and when running the scraper from `code/` it is written to `code/data/who_tb_global.csv`

3.4 Notes for reproducibility

- R version used for development: R 4.3.2 (scripts should work on R >= 4.0).
- The `scraping.R` script documents which WHO CSVs are used and contains the WDI code (indicator lists and query years).
- (Optional) Update and run ‘scraping.R‘ to refresh the merged dataset or directly go to next step. Scraped, combined, final dataset is available in the files uploaded.
- Ensure ‘app/data/who_tb_global.csv‘ exists (or re-generate with ‘scraping.R‘).
- Run ‘shiny::runApp("app")‘.

4 Methodology

- **Country-specific and global trends** were examined across six key TB indicators i.e. incidence, mortality, notifications, MDR-TB (percentage new), MDR-TB (percentage retreatment), and total RR-TB cases. Summary statistics such as the maximum, minimum, and most recent values were presented through interactive value boxes.
  - **Significance:** Tracking these indicators revealed regions showing improvement or decline and highlighted disruptions caused by the COVID-19 pandemic.
  - **Libraries used:** dplyr, ggplot2, plotly
- **Associations between TB burden and socioeconomic variables** such as GDP, poverty rate, and urbanization were explored using interactive scatter and line plots.
  - **Significance:** Since TB is strongly influenced by social and economic factors, visualizing these relationships provided valuable context to understand its distribution and to design targeted interventions.
  - **Libraries used:** dplyr, tidyr, ggplot2, plotly
- **Trends in drug-resistant TB (MDR-TB and RR-TB)** were analyzed to understand resistance patterns over time across countries.
  - **Significance:** Monitoring resistance supports the design of stronger treatment strategies and helps prevent further spread.
  - **Libraries used:** dplyr, ggplot2, plotly
- **The impact of COVID-19** was evaluated by comparing estimated TB incidence with reported cases to assess underreporting.
  - **Significance:** Pandemic-related disruptions affected TB detection and reporting, making this analysis vital for shaping recovery strategies.
  - **Libraries used:** dplyr, tidyr, ggplot2, plotly, viridis

5 Key Findings

5.1 Global Findings over the years

In this section we present our findings with graphs or plots made by the Shiny app.

1. **General Trend in countries over the years:** The following animated plots show the general incidence and mortality trends in top 10 countries over the period of 2000-2023. This showed that most of the affected countries in the recent years were African countries.

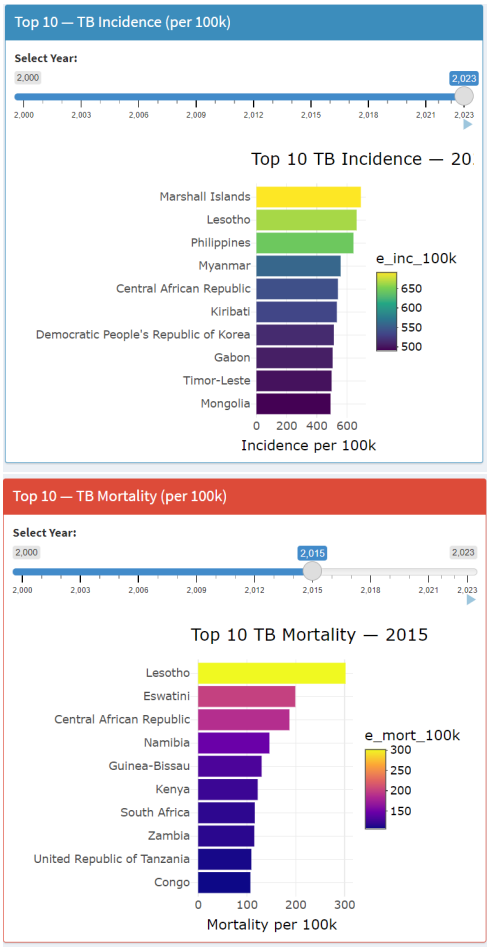

Figure 1: Home Page Animated Plots

2. **MDR-TB Trend:** The following plot shows the rapid decrease in MDR TB over the years, minimum being 2020, but sudden increase after that. This suggests interference of healthcare services due to Covid-19

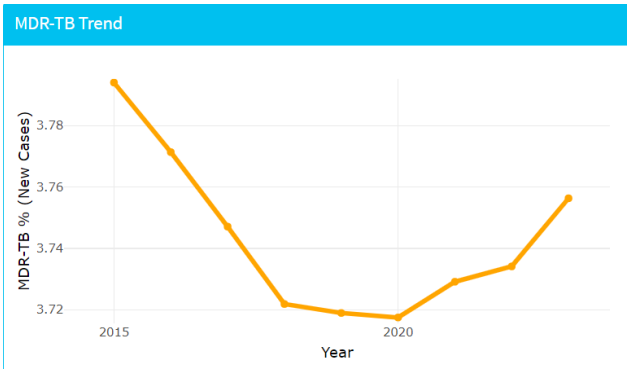

Figure 2: MDR-TB Graph

3. **Animated Bubble plots for Socioeconomic factors plotted with Incidences and Mortality:** These help to visualize which factors affected more over the years and the correlation between them. More can be tried on the app.

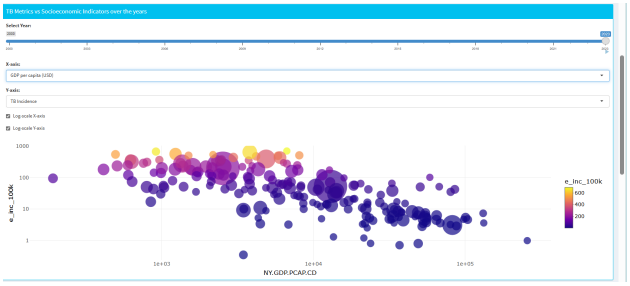

Figure 3: Socioeconomic factors' bubble plot

4. **Comorbidities**

   - **HIV Prevalence:** Measured Correlation between HIV prevalence and Incidences is 0.71. As HIV prevalence increases, the TB Incidence generally increases significantly. Countries with very low HIV prevalence tend to have low to moderate TB incidence. As HIV prevalence rises above 5-10%, TB incidence climbs dramatically, frequently reaching above 100 cases per 100k population and into the high-incidence (yellow/orange color) range, confirming that HIV is a major risk factor for TB.

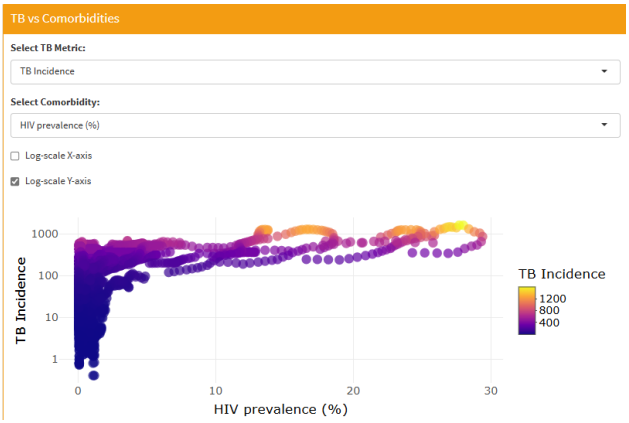

Figure 4: HIV vs Incidences

   - **Diabetes:** There is a weak or no clear correlation. The data points are scattered broadly across the plot.

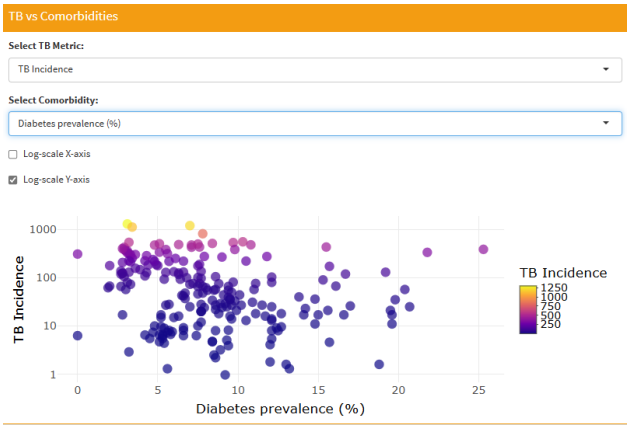

Figure 5: Diabetes vs Incidences

   - **Smoking:** There is a weak to moderate positive correlation. According to the statistical correlation measured, it is 0.04, which is insignificant.

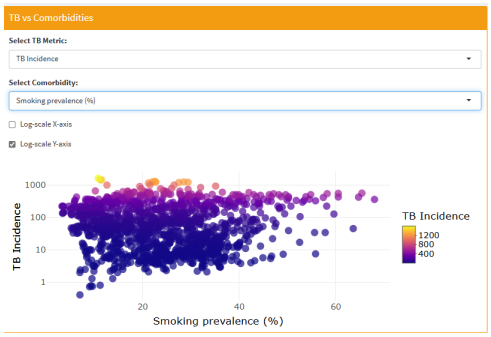

Figure 6: Smoking vs Incidences

5. **Environmental Factors- PM2.5 mean exposure:** High population exposure to poor

air quality (PM2.5) is a strong co-factor associated with the worst national TB epidemics. While low exposure doesn't guarantee low TB, widespread high exposure appears necessary for the highest incidence rates. It is not a perfect factor to see for correlation but it does affect.

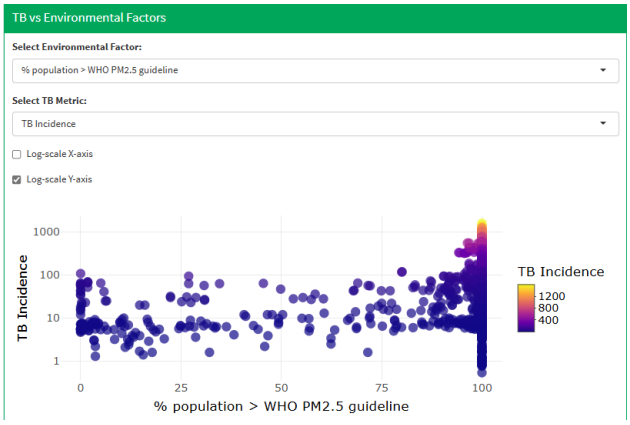

Figure 7: PM2.5 mean exposure vs Incidences

5.2 Domain specific trend Analysis

1. **TB Trends: Incidences and Mortality comparisons between countries**
   The line graphs plot trends of TB cases (incidence/morbidity) based on various parameters for the South Asian countries. However, in the app we can select any other country and any other parameter which we want to see. This multi-selection helps to compare regions such as African vs American while allowing single country statistics to be visible.

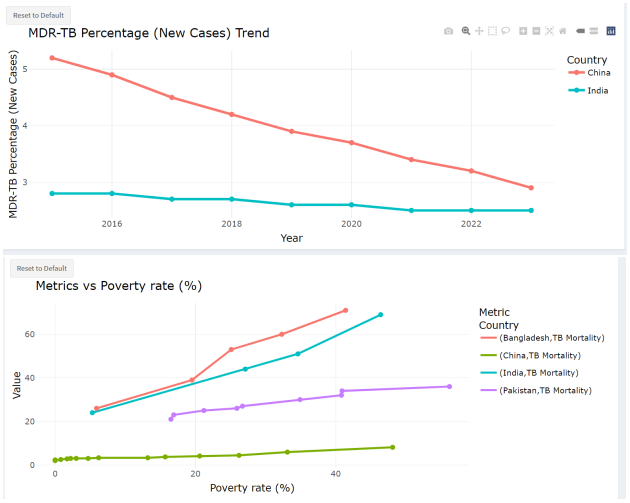

Figure 8: Different line graphs for trend analysis

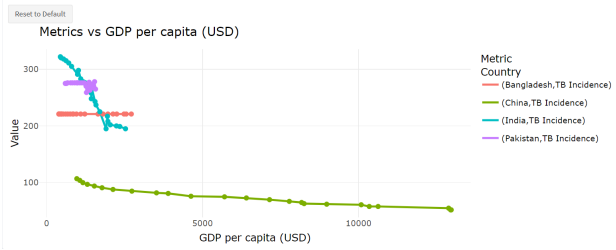

Figure 9: Different line graphs for trend analysis

2. **Socioeconomic Comparison:** This can be done for various TB related factors for various countries at once for efficient analysis. For example, we'll analyse India and China.

   Burden: At comparable GDP levels (e.g.$1,000 to $2,000 USD), India's TB Incidence was historically much higher than China's.

   Detection: Both countries show a convergence where Notifications follow Incidence, indicating improved detection over time, particularly as GDP rises. China's notifications consistently track its incidence more closely than India's.

Figure 10: Socioeconomic comparison plot

3. **MDR-TB Trends in countries:** MDR-TB cases are rising in some regions, especially where treatment adherence is low and high MDR-TB percentages are observed in countries with historically high TB incidence and weaker healthcare infrastructure.

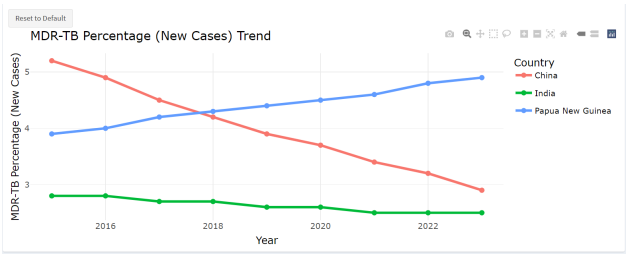

Figure 11: MDR TB trend plots

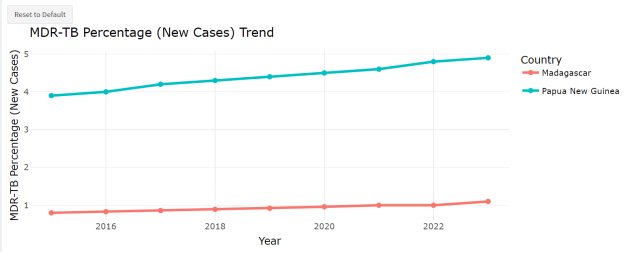

Figure 12: MDR TB trend plots

4. **Covid-19 Effect**: This comparison can be performed for any number of countries. For analysis, we have done for India and China.

   TB Burden: India had a consistently much higher estimated TB incidence (declining from ∼325 to ∼200 per 100k) than China (stable and low, declining from ∼100 to ∼50 per 100k), with both countries showing an overall decline.

   Detection Gap: India initially had a significant gap between estimated incidence and notifications (high underreporting), which rapidly closed post-2017, whereas China maintained low and stable underreporting throughout the period.

   COVID-19 Effect: The pandemic caused a severe, sharp drop in TB notifications and a corresponding peak in underreporting for India in 2020 (falling ∼50%), while China experienced only a minor dip in notifications.

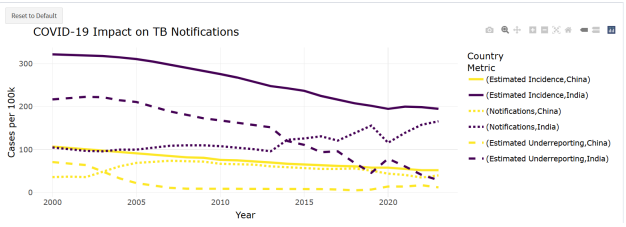

Figure 13: Covid Impact comparison

5. **Statistical Summary and Comparison:** Can be performed between any numeber of countries. For practical purposes, we have analysed for India and China.

   Density Comparison:
   1. China (Red): The density curve is narrow, sharply peaked, and centered around 75 cases per 100k, indicating that China's incidence rate has been stable and consistently low.

   2. India (Blue): The density curve is much broader and centered significantly higher, around 260 cases per 100k, indicating a higher and more variable historical burden.

   Boxplot Comparison:
   1. China: The boxplot is low on the Y-axis (around 70–90), confirming a consistently low burden with a small interquartile range.

   2. India: The boxplot is high on the Y-axis (around 200–300), visually representing the much higher incidence rate. The taller box confirms the higher historical variability (as noted by the larger SD) compared to China.

   In short, India has carried a much heavier and more variable TB burden than China between 2000 and 2023.

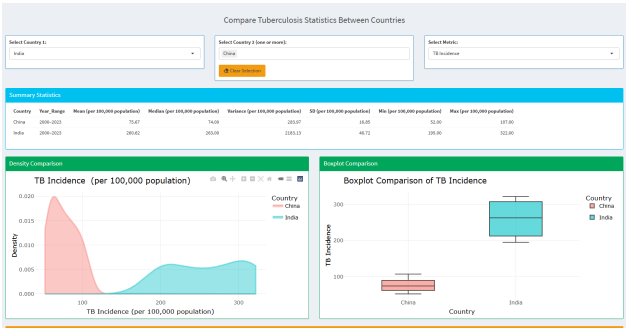

Figure 14: Statistical Comparison

6. **Correlation Heatmap between factors:** This heatmap shows an overall comparison among all factors influencing TB incidence and TB mortality amongst other factors. This helps us to realize which socioeconomic and environmental factors affect TB related factors more.

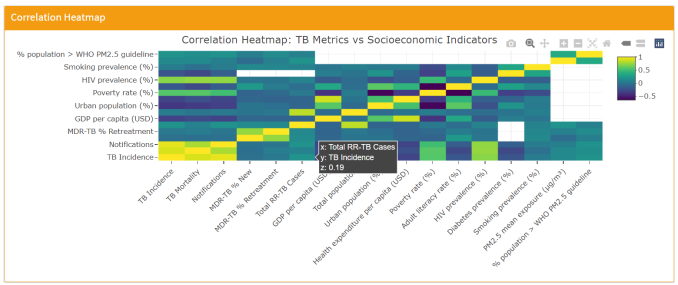

Figure 15: Correlation Heatmap

6 Conclusion

The project provides a statistical and visual analysis of global and country-specific TB patterns. Overall, the project contributes meaningfully to the understanding and monitoring of TB epidemiology and promotes data-driven strategies for global TB control. This can be extremely useful for research and educational purposes and help to monitor TB burden.

7 Limitations

- Current analysis is descriptive, lacks predictive modeling.
- Missing or incomplete WHO or WDI data introduce uncertainty in trend estimation.
- Extremely handpicked indicators, cannot be directly used on another WHO dataset since the variables are named differently

8 Future Improvements

- Apply ARIMA or exponential smoothing for TB incidence forecasting.
- Incorporate machine learning (e.g., Random Forest, XGBoost) for risk prediction.
- Add data imputation using statistical or ML techniques (e.g. KNN)

9 References

- WHO Global TB Programme Data
- World Bank World Development Indicators and WDI library
- Global Tuberculosis Report 2024
- The Social Determinants of Tuberculosis: From Evidence to Action
- Sociodemographic factors affecting knowledge levels of tuberculosis patients in New Delhi