HMS_520_TB_Final

Sophie Whikehart and Ye Htet Naing

2024-12-06

```
knitr::opts_chunk$set(echo = TRUE, message = FALSE)

# install libraries
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
```

Final Project - Global Mortality and Risk Factor Contributions for Tuberculosis Estimates from 2015 and 2020

Abstract

The World Health Organization (WHO) End TB Strategy set ambitious targets to reduce tuberculosis (TB) mortality by 35% and incidence by 20% between 2015 and 2020. While global progress has been reported, age-specific and regional analyses remain limited. This study utilizes the Global Burden of Disease 2021 (GBD 2021) Tuberculosis Estimates to assess trends in TB mortality and the contributions of smoking, alcohol use, and diabetes as risk factors.

Our analysis examines TB mortality trends from 2015 to 2020 across age groups, regions, and risk factors using mortality counts from the GBD data set.

Introduction

Tuberculosis (TB) is an ancient infectious disease with evidence of bone in TB in Egyptian mummies (2400 BC). TB is also known as consumption, the white plague, phthisis and the graveyard cough. TB is an airborne infection that is spread from someone with active TB. 1/10 people on average who are infected will go on to develop active TB. On the other hand, if you have latent TB you have no symptoms and cannot spread TB infection to others.

Testing for latent TB includes tuberculin skin tests and interferon gamma release assays [blood tests]. The sites of active TB disease include the lungs [pulmonary TB] and can also involve other organs such as the

bones and intestines. TB diagnosis is done with specimen collection of sputum, CSF or pleural fluid. The gold standard for diagnosis is through a culture but the disadvantage it that there is long turnaround time [up to 6-8 weeks]. Other methods include a smear microscopy and molecular methods such as Gene Xpert.

TB data sources often come from prevalence surveys, surveys measuring latent TB infection, TB case notifications, vital registration data and verbal autopsy data. However, data sources are imperfect and data quality and comparability can affect the measurement of the global burden of TB.

Tuberculosis (TB) continues to be one of the leading causes of infectious disease mortality globally, posing significant challenges to public health systems. In response, the World Health Organization (WHO) launched the End TB Strategy, aiming to reduce TB mortality by 35% and TB incidence by 20% between 2015 and 2020. While global progress has been documented, there is a limited understanding of how these trends vary by age and region, as well as the role of modifiable risk factors such as smoking, alcohol use, and diabetes. Age-specific evaluations are critical to inform tailored interventions and ensure equitable progress across populations.

In this study, we leverage the Global Burden of Disease 2021 (GBD 2021) Tuberculosis Estimates to examine TB mortality trends from 2015 to 2020 across different age groups and regions. The data set includes mortality counts stratified by location, year, and age, as well as estimates of deaths attributable to smoking, alcohol use, and diabetes.

Our primary research questions include:

1. How have mortality rates changed from 2015 to 2020 across different age groups and regions?
2. What is the relative contribution of different risk factors (e.g., smoking, alcohol use, and diabetes) to TB mortality in 2015 and 2020?
3. Do regions or age groups with higher reductions in TB mortality also show lower contributions of risk factors?

By addressing these questions, this project seeks to fill critical gaps in understanding TB mortality trends and their driving factors. Our findings will provide evidence to guide targeted interventions and strengthen global efforts toward achieving the WHO End TB Strategy goals.

Data Description

For this project we are using the data set from IHME titled: Global Burden of Disease 2021 [GBD 2021] Tuberculosis Estimates 1990-2021

This data set includes estimates of burden associated with all-form tuberculosis for GBD countries between 1990 and 2021. Tuberculosis mortality was informed by vital registration, verbal autopsy, sample-based vital registration and mortality surveillance data. TB morbidity data includes annual case notifications, data from prevalence surveys, and estimated cause specific mortality [CSMR] of TB among HIV-positive and HIV-negative individuals (IHME GBD 2021).

For our project we are utilizing the `IHME_GBD_2021_TB_MORTALITY_RISK_Y2024M03D19.XLSX` which contains risk deleted deaths due to all-form tuberculosis for alcohol use, smoking, and diabetes and all three risk factors combined by adult age groups by country for 2015, 2020 and 2021.

Methods

The variables of interest in our project were the mortality counts, risk factors and geographic and age group stratification. We pre-processed the data by:

1. Importing data set and filtering relevant columns
2. Checking and address missing data
3. Creating calculated variables

- Percent change in mortality between 2015 and 2020

- Proportional contribution of each risk factor to attributable TB deaths

We also conducted the following analyses

1. Trend Analysis

- Calculate percent change in mortality by region and age group from 2015 - 2020
- Visualize the trend using bar plot and line plot to highlight the change in mortality

2. Attributable Risk Factor Analysis

- Summarize the contribution of smoking, alcohol use and diabetes to TB mortality across age groups and regions
- Use grouped bar plots to visualize attributable mortality by risk factor

3. Association Between Mortality Reduction and Risk Factor Contribution

- Assess relationship between reduction in mortality rate and average risk factor contribution using linear regression model
- Visualize the association with scatter plot and regression line

```
# read in the data
mortality_data <- read_excel("data/IHME_GBD_2021_TB_MORTALITY_RISK_Y2024M03D19.XLSX")

## filter for relevant columns
mortality_data <- mortality_data %>%
  select(
    location_name,
    location_type,
    age_group_name,
    mort_2015_count_mean,
    mort_2020_count_mean,
    rmv_mean_smoking,
    rmv_mean_alcohol,
    rmv_mean_diabetes,
    rmv_mean_all_risk
  )

## check for missing values
colSums(is.na(mortality_data))
```

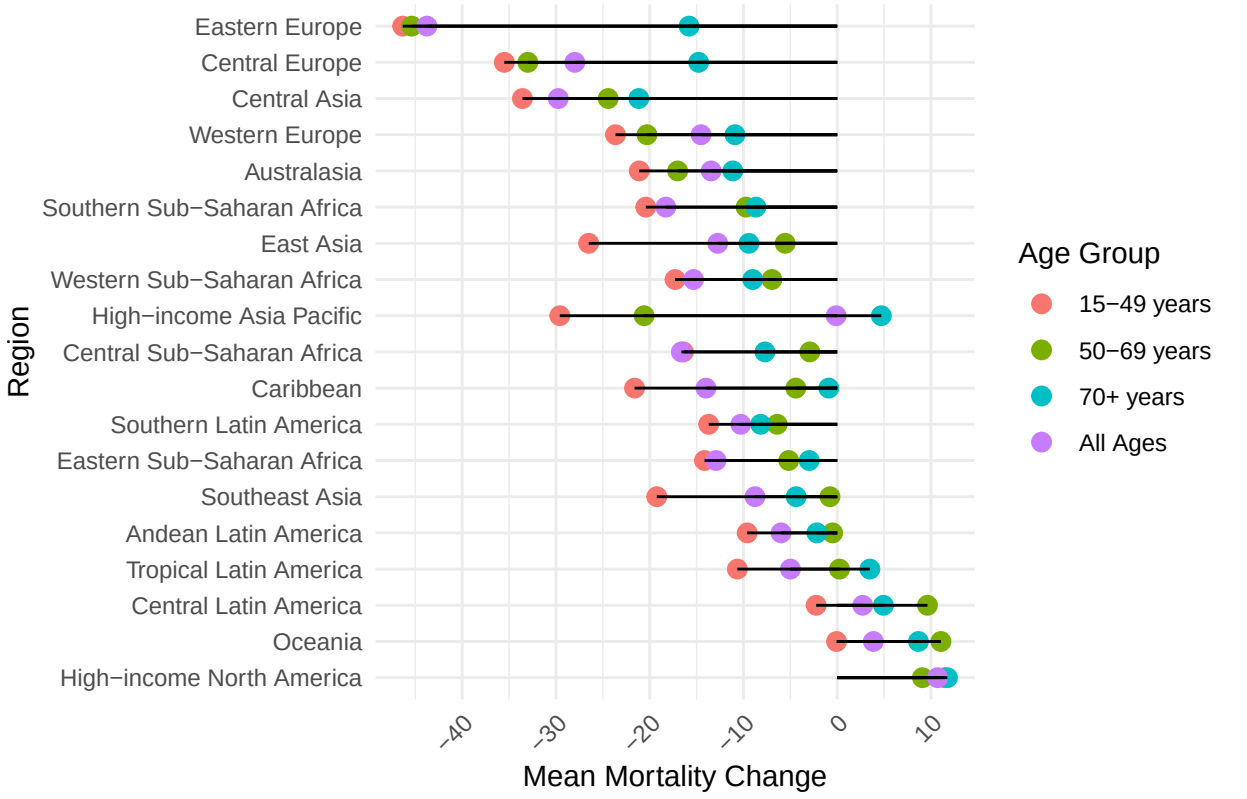
```
##        location_name        location_type       age_group_name
##                    0                    0                    0
## mort_2015_count_mean mort_2020_count_mean     rmv_mean_smoking
##                    0                    0                    0
##     rmv_mean_alcohol    rmv_mean_diabetes    rmv_mean_all_risk
##                    0                    0                    0
```

Question 1. How have mortality rates changed from 2015 to 2020 across different age groups and regions?

```
# filter for data just on region and super regions

region_data <- mortality_data %>%
  filter(location_type %in% c("region"))

# data wrangling -- calculate percent change in mortality by regions
```

```r
mortality_region <- region_data %>%
  mutate(mortality_change = ((mort_2020_count_mean - mort_2015_count_mean) / mort_2015_count_mean) * 100
  group_by(age_group_name, location_name) %>%
  summarize(mean_mortality_change = mean(mortality_change, na.rm = TRUE))

# visualization on region based on decreasing in mortality

## highlighting the largest reduction

fig_1 <- ggplot(mortality_region,
                aes(x = reorder(location_name, -mean_mortality_change),
                    y = mean_mortality_change,
                    color = age_group_name)) +  # Map color to age_group
  # Add points with variable colors
  geom_point(size = 3) +
  # Add lines from points to zero
  geom_segment(aes(x = location_name, xend = location_name, y = 0, yend = mean_mortality_change),
               color = "black") +
  # Flip axes for horizontal bars
  coord_flip() +
  # Add labels and title
  labs(
    title = "Fig 1. Mean Mortality Change from 2015 - 2020 (All Ages)",
    x = "Region",
    y = "Mean Mortality Change",
    color = "Age Group"  # Add legend title for color
  ) +
  # Minimal theme
  theme_minimal() +
  # Adjust axis text for readability
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )

fig_1
```

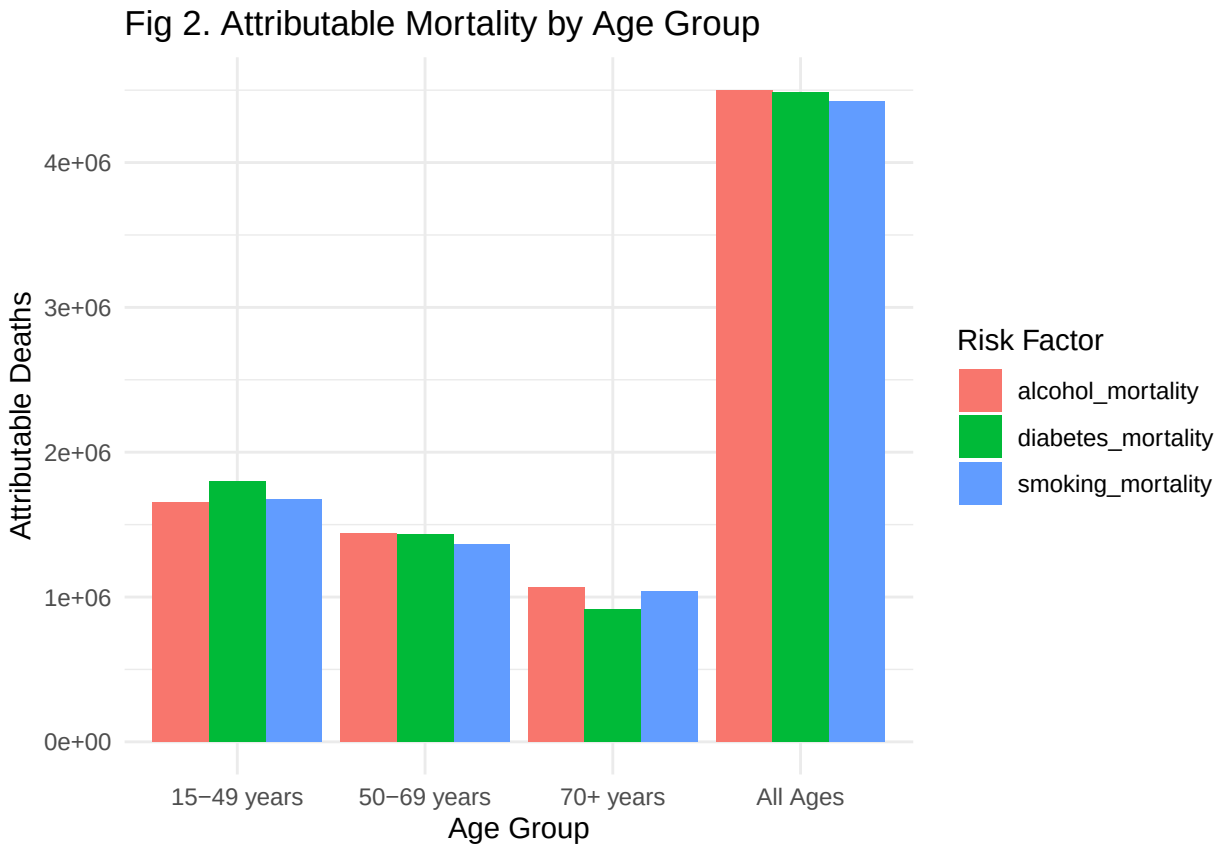
Fig 1. Mean Mortality Change from 2015 − 2020 (All Age



Question 2. What is the relative contribution of different risk factors [ex - smoking, alcohol, and diabetes] to mortality?

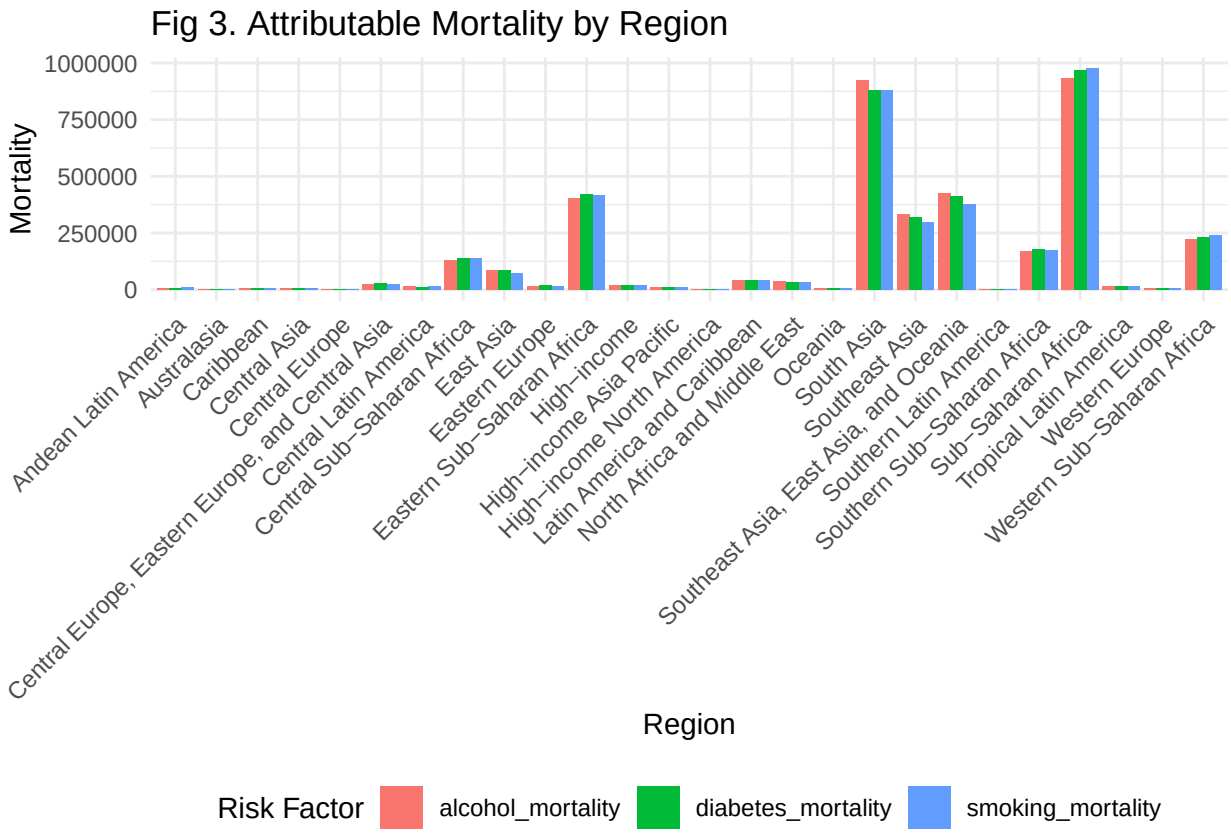
```
# Calculate attributable mortality

risk_factors <- mortality_data %>%
  mutate(
    prop_smoking = rmv_mean_smoking / rmv_mean_all_risk,
    prop_alcohol = rmv_mean_alcohol / rmv_mean_all_risk,
    prop_diabetes = rmv_mean_diabetes / rmv_mean_all_risk
  )

## this computes proportion of total attributable mortality due to each risk factor

# summarize by age

summary_by_age <- risk_factors %>%
  group_by(age_group_name) %>%
  summarise(
    smoking_mortality = sum(rmv_mean_smoking, na.rm = TRUE),
    alcohol_mortality = sum(rmv_mean_alcohol, na.rm = TRUE),
    diabetes_mortality = sum(rmv_mean_diabetes, na.rm = TRUE),
    total_attributable = sum(rmv_mean_all_risk, na.rm = TRUE)
  )
```

```r
# summarize_by_region <- risk_factors %>%


# visualize

summary_long <- summary_by_age %>%
  pivot_longer(
    cols = c(smoking_mortality, alcohol_mortality, diabetes_mortality),
    names_to = "risk_factor",
    values_to = "mortality"
  )

fig_2 <- ggplot(summary_long, aes(x = age_group_name, y = mortality, fill = risk_factor)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(
    title = "Fig 2. Attributable Mortality by Age Group",
    x = "Age Group",
    y = "Attributable Deaths",
    fill = "Risk Factor"
  ) +
  theme_minimal()

fig_2
```




```r
summarize_by_region <- risk_factors %>%
  filter(location_type %in% c("superregion", "region")) %>%  # Filter for superregion and region
  group_by(location_name, location_type) %>%  # Group by region and type
```

```
  summarise(
    smoking_mortality = sum(rmv_mean_smoking, na.rm = TRUE),
    alcohol_mortality = sum(rmv_mean_alcohol, na.rm = TRUE),
    diabetes_mortality = sum(rmv_mean_diabetes, na.rm = TRUE),
    total_attributable = sum(rmv_mean_all_risk, na.rm = TRUE)
  )

# Reshape to long format
region_long <- summarize_by_region %>%
  pivot_longer(
    cols = c(smoking_mortality, alcohol_mortality, diabetes_mortality),
    names_to = "risk_factor",
    values_to = "mortality"
  )

fig_3 <- ggplot(region_long, aes(x = location_name, y = mortality, fill = risk_factor)) +
  geom_bar(stat = "identity", position = "dodge") +  # Grouped bars
  labs(
    title = "Fig 3. Attributable Mortality by Region",
    x = "Region",
    y = "Mortality",
    fill = "Risk Factor"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),  # Rotate x-axis labels for readability
    legend.position = "bottom"
  )

fig_3
```



Question 3. Do regions or age groups with higher mortality reductions also show lower risk factor contributions?

```
# calculate mortality reduction rate

mortality_data <- mortality_data %>%
  mutate (
    reduction_rate = ((mort_2015_count_mean - mort_2020_count_mean) / mort_2015_count_mean) * 100
  )

# summary risk factor contribution

mortality_data <- mortality_data %>%
  rowwise() %>%
  mutate(
    avg_risk_factor_contribution = mean(c(rmv_mean_smoking, rmv_mean_alcohol, rmv_mean_diabetes), na.rm
  )

# linear regression

model <- lm(avg_risk_factor_contribution ~ reduction_rate, data = mortality_data)

# view model
summary(model)
```

```
##
```

```
## Call:
## lm(formula = avg_risk_factor_contribution ~ reduction_rate, data = mortality_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -10767   -9339   -8836   -7595 1226556
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9642.08    2308.68   4.176 3.24e-05 ***
## reduction_rate   -29.23     119.27  -0.245    0.806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56290 on 922 degrees of freedom
## Multiple R-squared:  6.513e-05,  Adjusted R-squared:  -0.001019
## F-statistic: 0.06006 on 1 and 922 DF,  p-value: 0.8065
```

```r
# visualize
fig_4 <- ggplot(mortality_data, aes(x = reduction_rate, y = avg_risk_factor_contribution)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(
    title = "Fig 4. Linear Regression: Mortality Reduction vs. Risk Factor Contributions",
    x = "Mortality Reduction Rate (%)",
    y = "Average Risk Factor Contribution"
  ) +
  theme_minimal()

fig_4
```

Results

In our first question we calculated and visualized percent changes in TB mortality across regions and age groups. Our results show that there are differences in reductions and increases in TB mortality by region and age group. In **Figure 1.** we calculated the mean mortality change from 2015 - 2020 by all ages and grouped by region and super region. The results of this figure show the visualization of region based on decreasing in mortality. Therefore Eastern Europe, Central Europe and Central Asia have seen the biggest reduction in mortality from 2015 to 2020. There are likely the areas where health interventions, improvements in healthcare or other factors have led to a reduction in deaths due to TB. On the other hand regions such as High-income North America, Oceania and Central Latin America have shown the highest increase in mortality. This suggests that these areas are where mortality rates have increased could be due to potential challenges such as poor healthcare access, economic difficulty and other systemic issues. This graph helps identify what regions need additional focus in terms of health interventions or resources to help reduce TB mortality.

This figure also highlights that consistently we see an increase of mortality change by the age groups in the 70+ and 50-69 year age group. Because these age groups consistently exhibit higher mortality change across multiple locations it indicates that there is need for age-specific interventions for older populations in all regions.

Our second question wanted to provide insight into the relative contributions of different risk factors (smoking, alcohol and diabetes) to mortality both by age group as well as by region. In **Figure 2.** the bar plot shows how attributable mortality for each risk factor varies across different age groups. As we can see from the graph in ages 15 - 49 the risk factor for diabetes is the highest with smoking slightly higher than alcohol mortality. In the age group of 50 - 69 alcohol is now the highest with diabetes second and smoking third. In 70+ age group, alcohol remains the highest with smoking second and diabetes third. Finally, in all ages alcohol is the highest with diabetes in second and smoking mortality third.

Therefore in ages 15-49 with diabetes being the highest contribute there could be a need for early screening and prevention programs for diabetes and promoting lifestyle modifications among younger populations. For ages 50-69 with alcohol being the leading risk factor in this age group interventions could include strengthening alcohol education and harm reduction initiatives as well as enhancing access to treatment for alcohol dependency. For ages 70+ alcohol remains the highest contributor which suggests long-term impact of alcohol use or age related vulnerabilities but smoking is the second highest so interventions could include the need for lifelong smoking cessation programs and addressing comorbidities related to alcohol and smoking in older adults.

Figure 3 shows the attributable mortality risk by region to see which regions have the highest morbidity due to different risk factors. From this figure we can see that regions with highest risk factor attributable deaths are Sub-Saharan Africa, South Asia and Eastern Sub-Saharan Africa. These likely reflect broader challenges such as limited access to healthcare and preventative services, socioeconomic inequalities that exacerbate exposure to risk factors and insufficient infrastructure for managing chronic diseases and addiction. In order to reduce these risk factors, public health policy and intervention priorities should be placed on these regions.

Finally our last question wanted to see if regions or ages groups with higher mortality reductions also showed lower risk factor contributions? We calculated the mortality reduction rate and then ran a linear regression model to provide insights about the relationship between mortality reduction and average risk factor contribution. The slope B1 was at -25.88 which indicates a very slight and non-significant decrease in average risk factor contribution for each 1% increase in mortality reduction rate. The p-value for the slop was 0.82 which indicates that there is no statistically significant relationship between mortality reduction and risk factor contribution. In addition the R^2 value was at 5.629 X 10^-5 which indicates that the reduction rate explains virtually none of the variation in the average risk factor contribution. The residuals also range quite widely to show that the data points are spread out and the model does not capture much variation.

In **Figure 4.** we visualized the results of our linear regression model and you can clearly see that there is no correlation. Because there is no significant relationship it is most likely that mortality reductions may be driven by other factors such as improvements in healthcare systems or better access to treatments rather than direct reductions in risk factor contributions. This highlights the need for broader exploration of mortality determinants beyond the examined risk factors.

Conclusion

This project aimed to assess the trends in tuberculosis mortality from 2015 - 2020 and examine the contribution of smoking, alcohol use and diabetes as risk factors. Our analysis revealed several important findings

1. TB Mortality Trends
2. Risk Factor Contribution
3. Association between Mortality Reduction and Risk Factor Contribution

These findings show the importance of addressing both the disease itself and modifiable risk factors contributing to TB mortality. Targeted interventions based on the findings from this report could potentially further accelerate progress toward the WHO End TB Strategy's mortality reduction goals. Additionally region and age specific approaches tailored to local health challenges are crucial to achieve sustained reductions in TB mortality globally.

```
# save figures

# Save the plot
ggsave(
  filename = "figures/fig_1.jpg",
  plot = fig_1
)

ggsave(
```

```
  filename = "figures/fig_2.jpg",
  plot = fig_2
)

ggsave(
  filename = "figures/fig_3.jpg",
  plot = fig_3
)

ggsave(
  filename = "figures/fig_4.jpg",
  plot = fig_4
)
```