hms_520_TB_Final

Sophie Whikehart and Ye Htet Naing

2024-12-06

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

Abstract

- Ye
- Give an overview of the project, what our goals are and what we hope to find

Introduction

- Ye

Provide a brief background: what do we already know about this topic? Why is it interesting? What do you want to add to what already exists?

Data Explanation

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

Data pre-processing

```
# read in the data
mortality_data <- read_excel("/Users/seanwilcox/Desktop/HMS_520_Final_Project_TB/data/IHME_GBD_2021_TB_

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

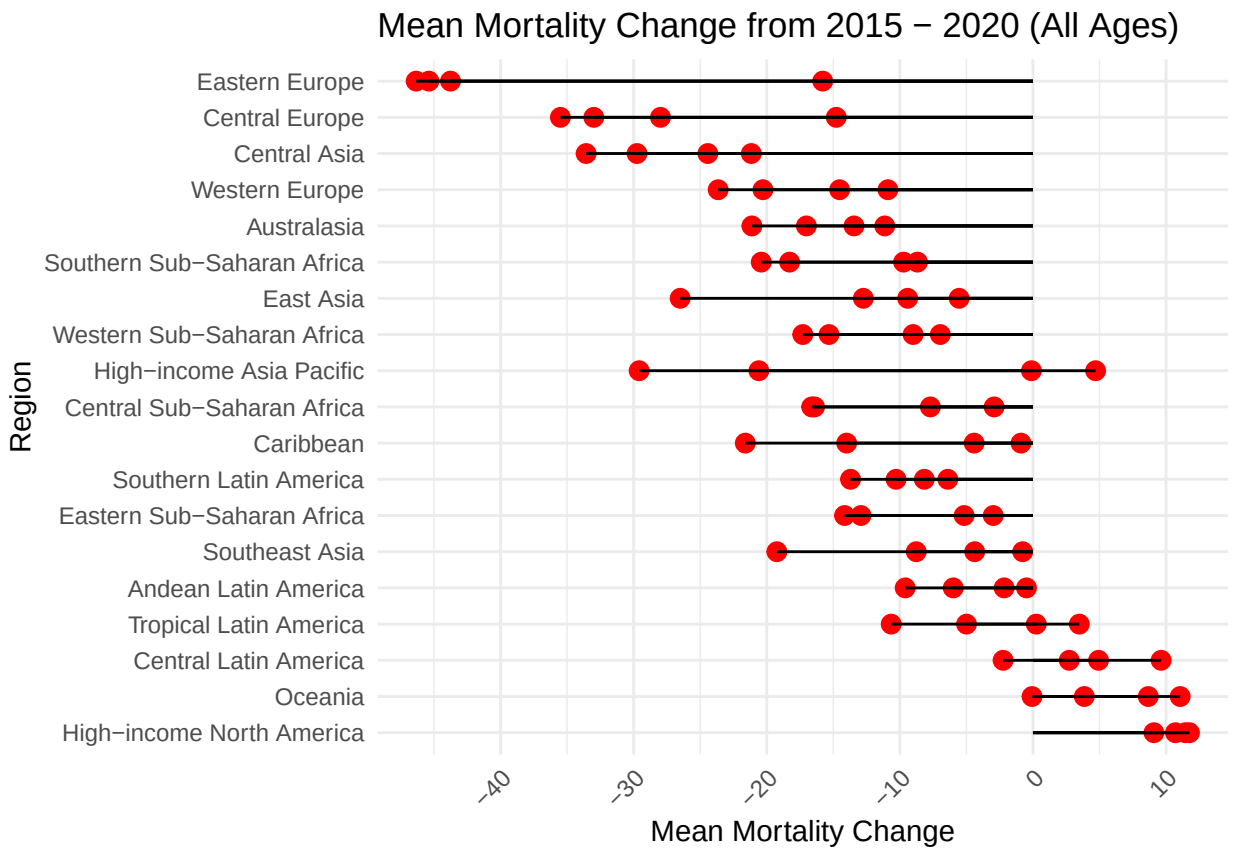
```
##         location_name         location_type        age_group_name
##                     0                     0                     0
##           location_id  mort_2015_count_mean mort_2015_count_lower
##                     0                     0                     0
## mort_2015_count_upper  mort_2020_count_mean mort_2020_count_lower
##                     0                     0                     0
## mort_2020_count_upper      rmv_mean_smoking     rmv_lower_smoking
##                     0                     0                     0
##     rmv_upper_smoking      rmv_mean_alcohol     rmv_lower_alcohol
##                     0                     0                     0
##     rmv_upper_alcohol     rmv_mean_diabetes    rmv_lower_diabetes
##                     0                     0                     0
##    rmv_upper_diabetes     rmv_mean_all_risk    rmv_lower_all_risk
##                     0                     0                     0
##    rmv_upper_all_risk
##                     0
```

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

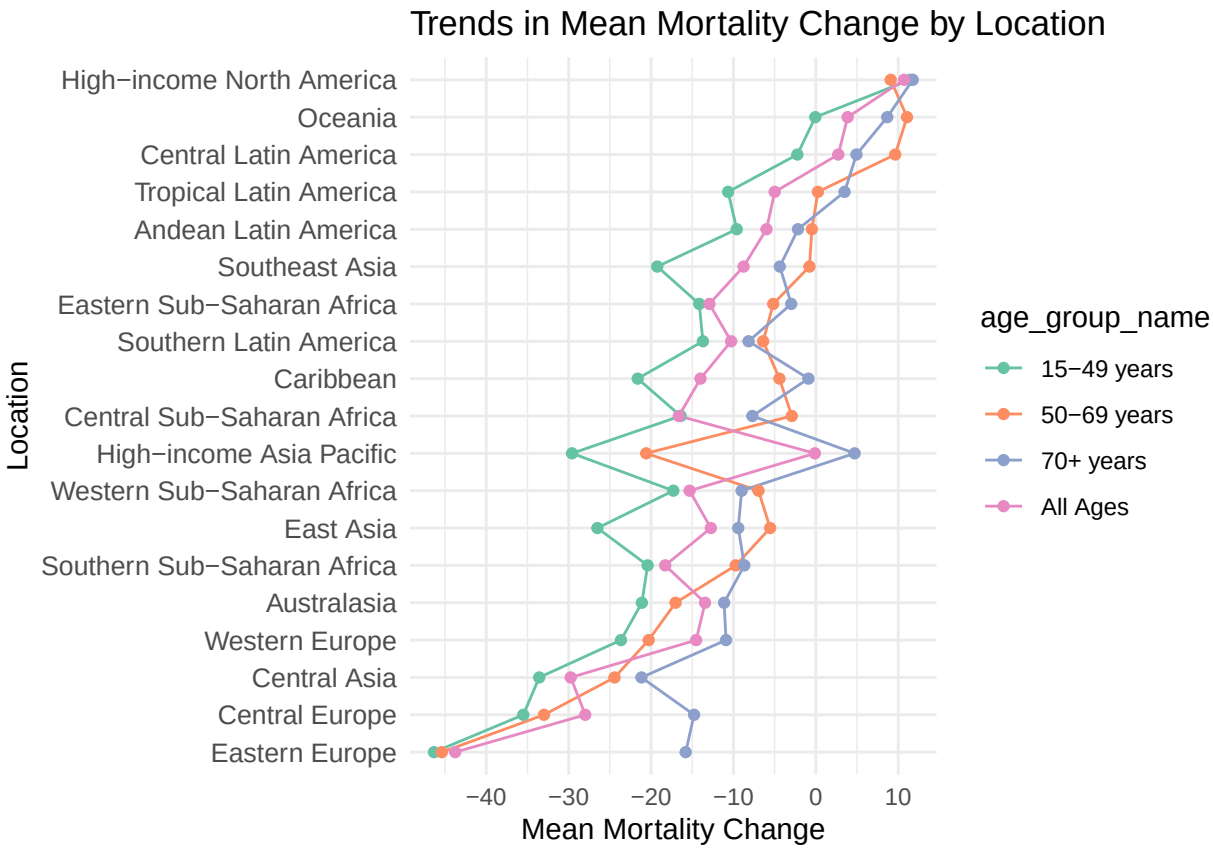
```
# filter for data just on region and super regions

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

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

mortality_region <- region_data %>%
  mutate(mortality_change = ((mort_2020_count_mean - mort_2015_count_mean) / mort_2015_count_mean) * 10
  group_by(age_group_name, location_name) %>%
  summarize(mean_mortality_change = mean(mortality_change, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'age_group_name'. You can override using
## the `.groups` argument.
```

```
# visualization on region based on decreasing in mortality

## highlighting the largest reduction

ggplot(mortality_region, aes(x = reorder(location_name, -mean_mortality_change), y = mean_mortality_cha
  geom_point(color = "red", size = 3) +
  geom_segment(aes(xend = location_name, yend = 0), color = "black") +  # Lines from points to zero
  coord_flip() +  # Flip axes for horizontal bars
  labs(title = "Mean Mortality Change from 2015 - 2020 (All Ages)",
       x = "Region",
       y = "Mean Mortality Change") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
## Regions with the largest negative values (located at the top of the plot) have seen the biggest redu

## Regions with positive values will show increased mortality, suggesting areas where the mortality rat

## plot provides a comparative view of how different regions have performed in terms of mortality over

## If you are analyzing regional disparities, the plot helps identify which regions have been improving

# Create a line plot to visualize trends by age group
ggplot(mortality_region, aes(x = reorder(location_name, mean_mortality_change),
                             y = mean_mortality_change,
                             color = age_group_name,
                             group = age_group_name)) +
  geom_line() +  # Adds lines for each age group
  geom_point() +  # Adds points at each location for clarity
  coord_flip() +  # Flip coordinates for horizontal lines
  labs(title = "Trends in Mean Mortality Change by Location",
       x = "Location",
       y = "Mean Mortality Change") +
  theme_minimal() +  # Clean minimal theme
  theme(axis.text.y = element_text(size = 10)) +  # Adjust text size for readability
  scale_color_brewer(palette = "Set2")  # Use a color palette for age groups
```




```
## represents average change in mortality rates for each location
  ## positive value indicates increase in mortality while negative value indicates decrease

## lines and points corresponding to each age group show how mean mortality change trend varies for tha

## slope of each line indicates direction and strength of change
  ## if line is sloped upward it indicates that mortality has increased in those locations for that age
```

2. What is the relative contribution of different risk factors [ex - smoking, alcohol, and diabetes] to mortality in 2015 and 2020?
3. Do regions or age groups with higher mortality reductions also show lower risk factor contributions?

Results

- Ye & Sophie