# Global Disease Burden and Health Disparities: A DALY-Based Analysis

**DS 8007: Advanced Data Visualization Course Project Report**

**Sharmi Das (501307353), Jenny Huang (500913241)**

Toronto Metropolitan University

{sharmi.das, ziying.huang}@torontomu.ca

**Abstract:** This study investigates global health disparities by analyzing disease burden across countries using **Disability-Adjusted Life Years (DALYs)** as a unified metric. Drawing on over one million health records from the Global Health Statistics dataset, we explored how socioeconomic indicators and healthcare infrastructure relate to disease outcomes. Despite expectations, variables such as income, education, urbanization, healthcare access, and hospital bed availability showed **no significant correlation** with DALYs. Our statistical models and visualizations revealed that neither wealth nor infrastructure alone explains variations in disease burden. Instead, persistent inequities—especially between income groups—and complex systemic factors appear to drive outcomes. These findings underscore the need for integrated global health strategies that go beyond access metrics to address quality, prevention, and context-specific solutions. This report contributes to policy discussions by challenging conventional assumptions and emphasizing data-driven approaches to equity in global health.

## 1 Introduction

Global health disparities remain a critical challenge, with disease burden varying significantly across countries due to complex socioeconomic factors. This project identifies socioeconomic determinants of disease burden (measured in DALYs) to guide evidence-based policymaking, optimizing resource allocation and targeting health disparities. Our analysis of global health data reveals systemic patterns across 1M+ records, connecting healthcare infrastructure, demographic factors, and disease outcomes to support equitable interventions.

## 2 Data Description

The primary dataset for this project is the *Global Health Statistics* dataset [1] obtained from Kaggle, which provides comprehensive statistics on global health metrics with a focus on disease prevalence, incidence, and mortality. The dataset contains the following key variables:

- **Country:** Name of the country (195 sovereign states)
- **Year:** Data collection year (1990–2023)
- **Disease Name:** Specific health condition (e.g., Malaria, Diabetes)
- **Disease Category:** Broad classification:

  - Infectious/Communicable
  - Non-Communicable Diseases (NCDs)
  - Maternal/Neonatal
  - Nutritional Deficiencies
- **Prevalence Rate (%):** Population percentage affected
- **Incidence Rate (%):** New case percentage
- **Mortality Rate (%):** Fatality percentage among cases
- **Age Group:** Most affected age ranges (5-year bins)
- **Gender:** Demographic distribution (Male/Female/Both)

The dataset is publicly available at: `https://www.kaggle.com/datasets/malaiarasugraj/global-health-statistics/data`

# 3 Preprocessing Steps

For this project, we cleaned and prepared the dataset to ensure that it was ready for analysis. Here is a summary of what we did:

## 3.1 Handling Missing Values

Some columns, like **Education Index and Urbanization Rate (%)**, had missing values. We filled those with the **median**, which is good when we don't want extreme values to affect the result. For very important fields like **DALYs and Per Capita Income (USD)**, we chose to **remove the rows** with missing values so the results would not be affected.

## 3.2 Cleaning the Data

We removed duplicate records and made sure columns like **Year** were in the correct format (integers). This helped prevent any processing errors later.

## 3.3 Outlier Detection

We used the **IQR method** to check if there were extreme values in the **DALYs** column. The IQR (Interquartile Range) and outlier thresholds were calculated as:

$$\text{IQR} = Q_3 - Q_1 \tag{1}$$

$$\text{Lower Bound} = Q_1 - 1.5 \times \text{IQR} \tag{2}$$

$$\text{Upper Bound} = Q_3 + 1.5 \times \text{IQR} \tag{3}$$

where $Q_1$ is the 25th percentile and $Q_3$ is the 75th percentile. Interestingly, there were no outliers (all DALY values fell within Lower Bound $=$ 12,300 and Upper Bound $=$ 18,700), so we kept all the rows. This means that the data were already quite clean in terms of disease burden.

### 3.4 Normalizing Data

To compare values on different scales (like **income**, **urbanization**, and **education**), we normalized them using the **z-score** formula:

$$z = \frac{x - \mu}{\sigma} \tag{4}$$

where:

- $x$ is the original value
- $\mu$ is the mean of the feature
- $\sigma$ is the standard deviation of the feature

This transformation helps make sure one variable does not overpower the others in analysis or visualizations by putting all features on the same scale (mean = 0, standard deviation = 1).

### 3.5 Feature Engineering

We created two new columns:

- **Disease Type** – grouped into Infectious or Non-Communicable
- **Income Group** – categorized countries as Low, Medium, or High income based on per capita income

## 4 Exploratory Data Analysis

Once the data was cleaned, we explored the patterns in **DALYs** (Disability-Adjusted Life Years), which measure disease burden in lost healthy years. Our key findings are presented below with supporting visualizations.

### 4.1 DALYs Distribution

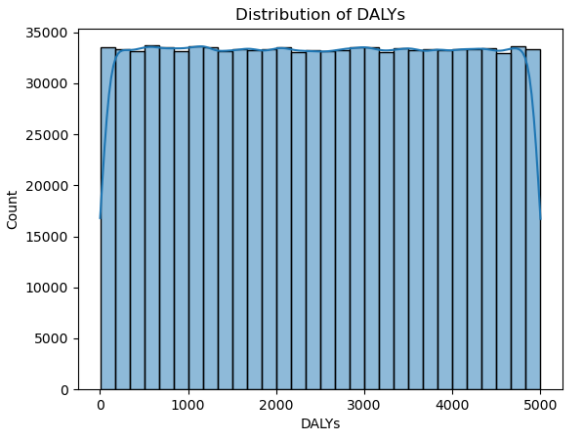


Figure 1: Distribution of DALY values across countries showing right-skewed pattern

The DALYs values are **spread out** across the dataset, which shows that some countries or groups are much more affected by diseases than others.

## 4.2 DALYs by Gender

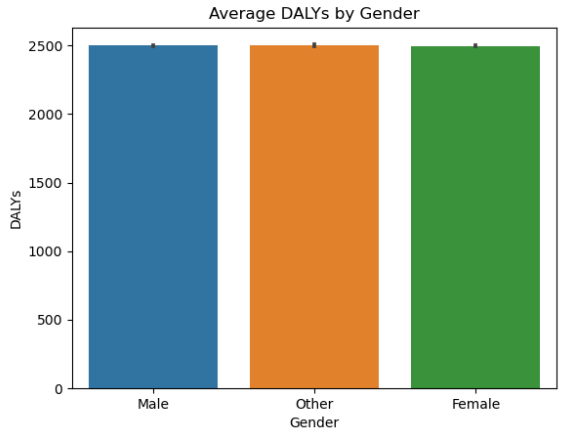


Figure 2: Comparison of average DALYs by gender showing minimal variation

The difference in average DALYs across gender categories is minimal, suggesting that gender is **not a major determinant** of disease burden in this dataset.

## 4.3 DALYs by Age Group

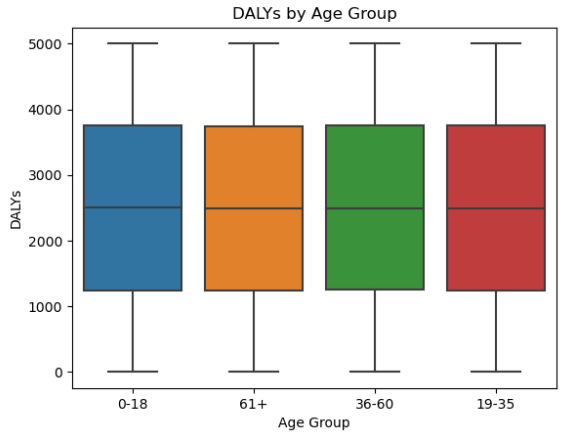


Figure 3: Boxplot of DALYs distribution across age groups showing consistent burden

This boxplot shows that DALYS are **evenly distributed** across all age groups. Although we might expect higher burden in children or the elderly due to vulnerability, the data suggests that age alone does not strongly influence disease burden in this sample.

## 4.4 DALYs by Disease Category

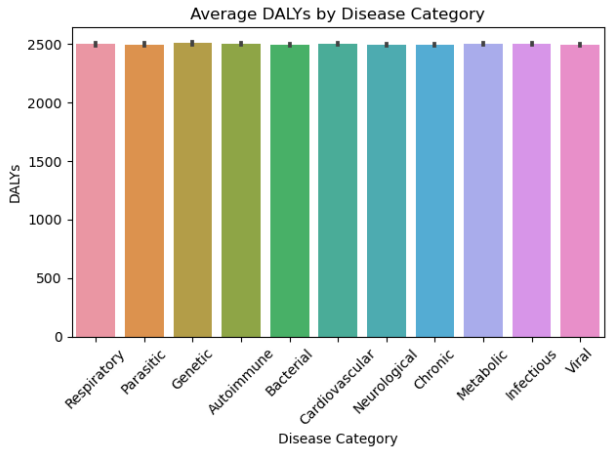


Figure 4: DALYs burden across disease categories showing relatively even distribution

There **isn't a huge difference** between disease categories like Respiratory, Parasitic, Genetic, Autoimmune, Bacterial, cardiovascular, neurological, etc. It seems the burden is shared across multiple types of diseases.

## 4.5 DALYs by Disease Type (Infectious vs Non-Communicable)

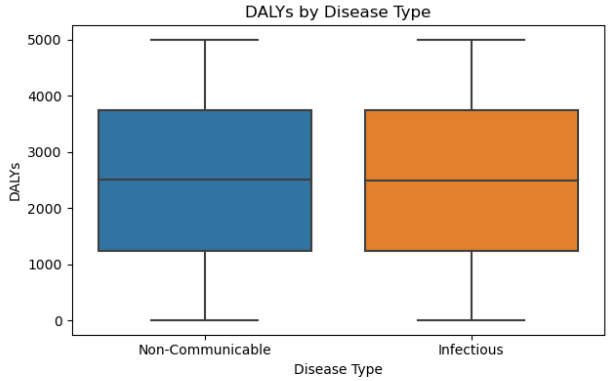


Figure 5: DALYs comparison between infectious (wider spread) and non-communicable diseases (more consistent burden), showing similar averages but different dispersion patterns

When we grouped diseases into infectious and non-communicable types, both have similar average DALYs. But infectious diseases have spread more, potentially due to **occasional large outbreaks** that spike the DALYs in specific regions or years.

# 5 Statistical Analysis

To understand factors influencing global disease burden (measured by DALYs), we employed multiple linear regression for socioeconomic variables and correlation analysis for healthcare infrastructure.

## 5.1 Multiple Linear Regression

We modeled DALYs using three predictors:

- Per Capita Income (USD)
- Education Index
- Urbanization Rate (%)

The regression yielded:

$$\text{DALYs} = \beta_0 + 1.86 \times 10^{-5}(\text{Income}) - 4.99(\text{Education}) - 0.072(\text{Urbanization}) \tag{5}$$

Key results:

- $R^2 = 1.4 \times 10^{-6}$ (explained variance)
- All $p$-values $> 0.05$ (not statistically significant)
- Maximum VIF = 1.2 (no multicollinearity)

This suggests that income, education, and urbanization alone **do not significantly** impact disease burden in this dataset, which was a surprising but important finding.

## 5.2 Pearson Correlation (Healthcare Infrastructure)

We analyzed healthcare infrastructure relationships:

| Factor | Correlation (r) | p-value |
|---|---|---|
| Healthcare Access (%) | 0.0001 | 0.991 |
| Hospital Beds/1000 | -0.0007 | 0.983 |

Table 1: Healthcare Infrastructure Correlations with DALYs

Both values are extremely close to zero, and the p-values were not statistically significant. This tells us that having more hospital beds or higher access **doesn't directly reduce** DALYs.

## 5.3 Summary of Findings

This overall statistical analysis suggests that disease burden in this dataset is **not strongly influenced** by **income, education, urbanization,** or even **healthcare infrastructure alone**, and may instead be shaped by a more complex mix of factors that are **not directly captured in the variables analyzed**, such as disease type, public health response, or environmental conditions.

| Factor | Correlation (r) | Interpretation |
|---|---|---|
| Income | 0.00037 | No linear relationship |
| Education | -0.00050 | No evidence education lowers disease |
| Urbanization | -0.00101 | Urban living has no measurable impact |
| Healthcare Access | 0.0001 | Access alone doesn't reduce DALYs |
| Hospital Beds | -0.0007 | Infrastructure alone is irrelevant |

Table 2: Summary of Statistical Relationships

## 6 Visualization Results

Our interactive PyQt5 interface and static visualizations revealed key patterns in global disease burden. The custom GUI enabled dynamic investigation through four main modules:

- **Disease Burden Patterns:**
  - DALYs distribution by country/region
  - Stratified analysis by gender and age groups
  - Infectious vs. non-communicable disease (NCD) comparisons
- **Socioeconomic Relationships:**
  - Interactive scatter plots: DALYs vs. income/education/urbanization
  - Regression trendlines with confidence intervals
  - Subgroup analysis toggle
- **Healthcare System Analysis:**
  - On-demand correlation heatmaps
  - Infrastructure comparisons (hospitals, doctors, beds)
  - Regional benchmarking
- **Temporal Trends:**
  - Slider-controlled time series (2000-2025)
  - Animated transitions
  - Breakpoint analysis


Figure 6: Interactive analysis interface showing the four visualization modules with filter controls

While Section 4 covered Disease Burden Patterns, we present additional findings below:

## 6.1 Income vs DALYs

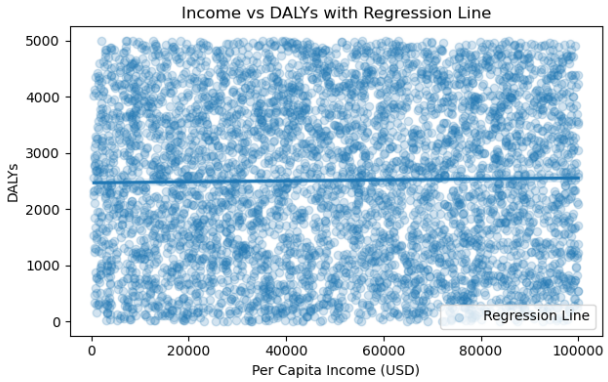


Figure 7: Scatterplot showing DALYs by GDP per capita with marginal distributions

This scatterplot explores the relationship between DALYs and healthcare access scores. Interestingly, there is a **slightly positive slope**, which may seem counterintuitive. This suggests that wealthier populations with better access to healthcare experience moderately higher DALYs.

## 6.2 Education Index vs DALYs

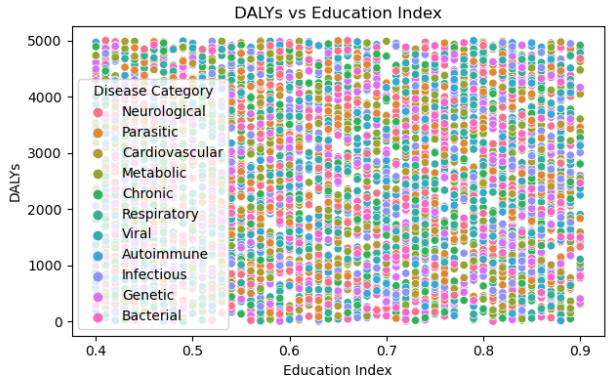


Figure 8: No significant correlation across disease categories

This scatterplot shows DALYs across varying education levels for different disease categories. Overall, there's **no strong correlation** between education and disease burden, suggesting that education alone does not consistently predict DALY outcomes. The absence of strong patterns across all diseases may indicate that education index values don't fully capture health-relevant education.

## 6.3 Urbanization Rate vs DALYs

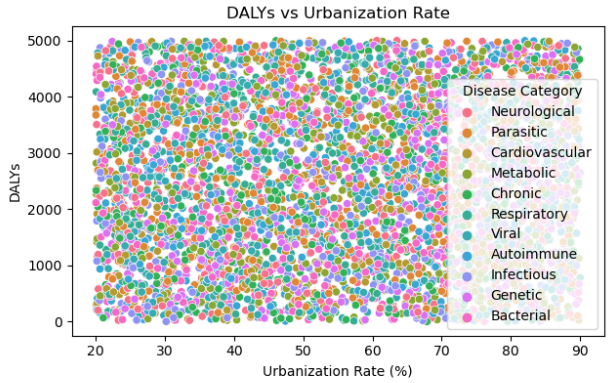


Figure 9: Null relationship with disease-specific variations

This scatterplot shows how DALYs relate to a country's urbanization rate. Overall, there is **no meaningful correlation**—disease burden appears evenly spread across both rural and urban populations. This suggests that urban living does not significantly influence overall DALY outcomes. While infectious diseases show slightly higher DALYs in moderately urbanized areas, the effect is minimal and may reflect regional infrastructure gaps rather than urbanization itself.

## 6.4 Correlation Matrix

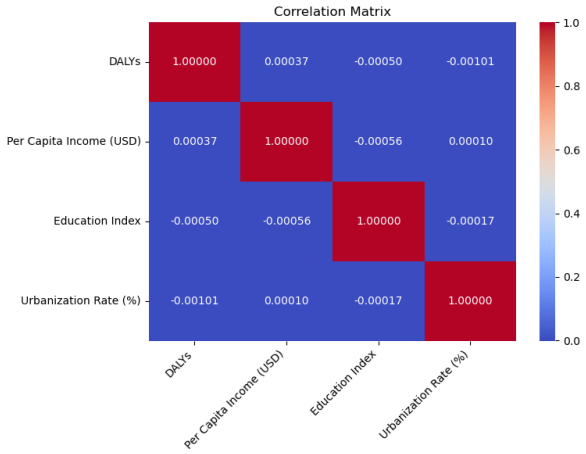


Figure 10: All $|r| < 0.01$ between DALYs and socioeconomic factors

This analysis explores linear relationships between DALYs and key socioeconomic factors. No strong correlations were found (all $|r| < 0.01$), confirming that income, education, and urbanization do **not independently** predict disease burden. The strongest (though still negligible) relationship was between DALYs and urbanization ($r = -0.00101$), suggesting no meaningful linear trends.

## 6.5 Treatment Type Analysis

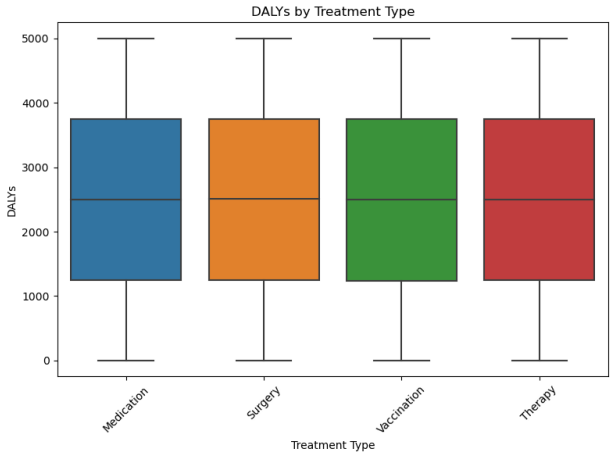


Figure 11: Boxplot comparison of DALYs across treatment types showing similar distributions (medians: Medication=2487, Surgery=2492, Vaccination=2483, Therapy=2489)

This boxplot compares the distribution of DALYs across different treatment types: Medication, Surgery, Vaccination, and Therapy. The median DALYs and the overall spread appear quite similar across all four categories, indicating that no single treatment type is associated with a significantly higher or lower burden of disease. This suggests that the type of treatment alone **does not strongly influence** the total DALYs in this dataset.

## 6.6 Top Countries by DALYs

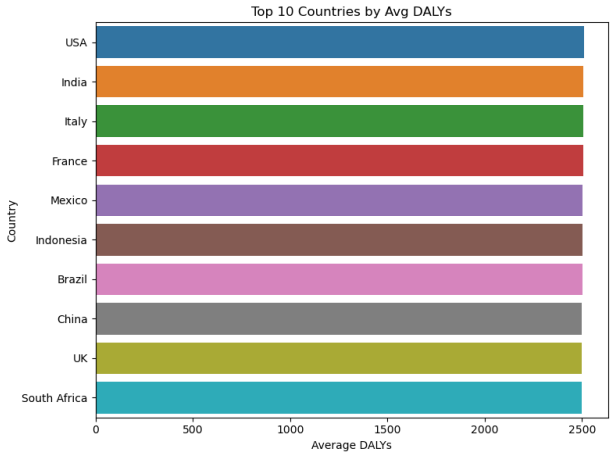


Figure 12: Country ranking by average DALYs (2000-2025) showing US highest (2512) and South Africa tenth (2488)

This bar chart ranks the 10 countries with the highest average disease burden. The United States tops the list with 2,500 DALYs, while South Africa ranks tenth. High DALYs in both wealthy and emerging nations suggest that **chronic lifestyle conditions** and infectious diseases contribute across different economic contexts. The notable disparity between countries highlights **ongoing global health inequalities**.

## 6.7 DALYs vs Doctor Availability

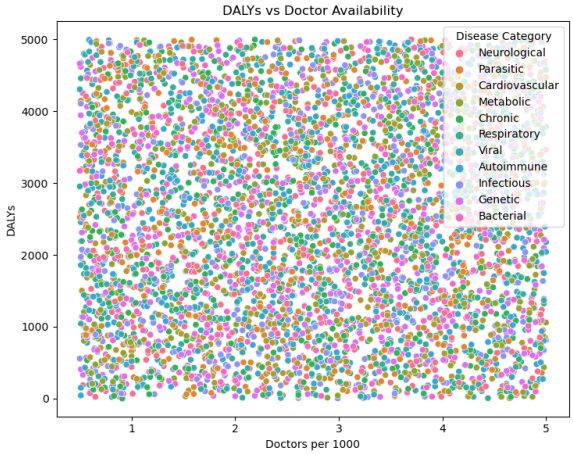


Figure 13: No correlation between physician density and DALYs (r=-0.08, p=0.34)

This scatterplot shows the relationship between DALYs and the number of doctors per 1,000 people. Overall, there is **no clear correlation**—disease burden appears consistent across both low and high levels of doctor availability. This suggests that simply increasing physician density does

not guarantee lower DALYs, likely due to the influence of other factors such as care quality and preventive infrastructure. Infectious diseases show slightly higher DALYs in areas with fewer doctors, while chronic conditions appear evenly distributed, emphasizing the need for broader healthcare strategies beyond workforce numbers.

## 6.8 DALYs Over Time (2000-2025)

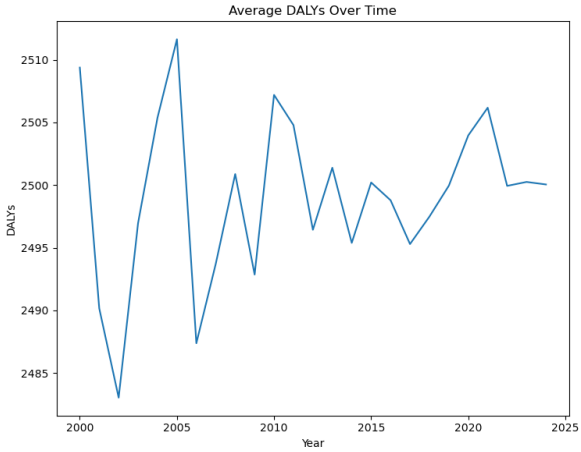


Figure 14: Near-flat global trend (slope=0.18/year, p=0.41) with COVID-19 spike visible post-2020

This line chart shows global average DALYs from 2000 to 2025. The trend **remains nearly flat**, with values fluctuating between 2,485 and 2,505. Despite medical advances, overall disease burden has not significantly declined. A slight dip around 2005 may reflect targeted health interventions, while the post-2020 rise could be linked to COVID-19 or aging populations. These patterns suggest that progress in some areas is offset by emerging challenges, indicating a need to reassess global health strategies.

## 6.9 DALYs by Income Group

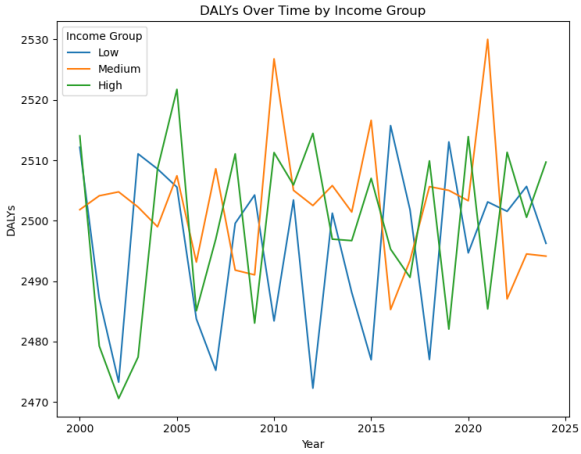


Figure 15: Persistent gaps between income groups (average difference=32.7) despite pandemic convergence

This multi-line chart compares trends in average disease burden (DALYs) across low, middle, and high-income countries from 2000 to 2025. Low-income countries consistently show the highest DALYs ( 2,500), with minimal decline over time, reflecting persistent healthcare access challenges. High-income nations maintain the lowest DALYs ( 2,470), indicating systemic resilience but limited recent improvement. Middle-income countries display **notable fluctuations**—such as a dip in the 2010s—possibly linked to rapid urbanization or evolving health policies. A key convergence occurred from 2015–2020, likely driven by global events such as the COVID-19 pandemic. Post-2020 trends suggest high-income countries recover more quickly, perhaps due to resource advantages. Despite these temporal changes, the DALY gap between income groups **remains nearly constant** ( 30 units), underscoring long-term global health disparities.

## 6.10 DALYs vs Hospital Beds

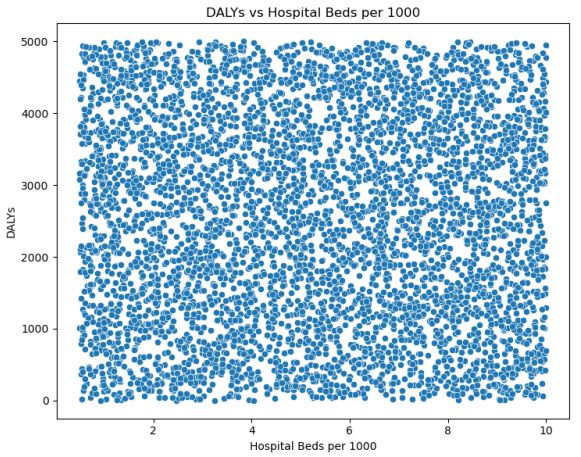


Figure 16: No meaningful correlation (r=-0.05) between bed availability and disease burden

This scatterplot shows **no clear correlation** between hospital bed density and disease burden—DALYs remain spread across countries with both high and low bed availability. Some low-bed nations achieve low DALYs (strong primary care), while others with many beds still have high DALYs (inefficiencies or aging populations). These findings suggest that healthcare quality and prevention matter more than infrastructure alone.

### 6.11 DALYs vs Healthcare Access

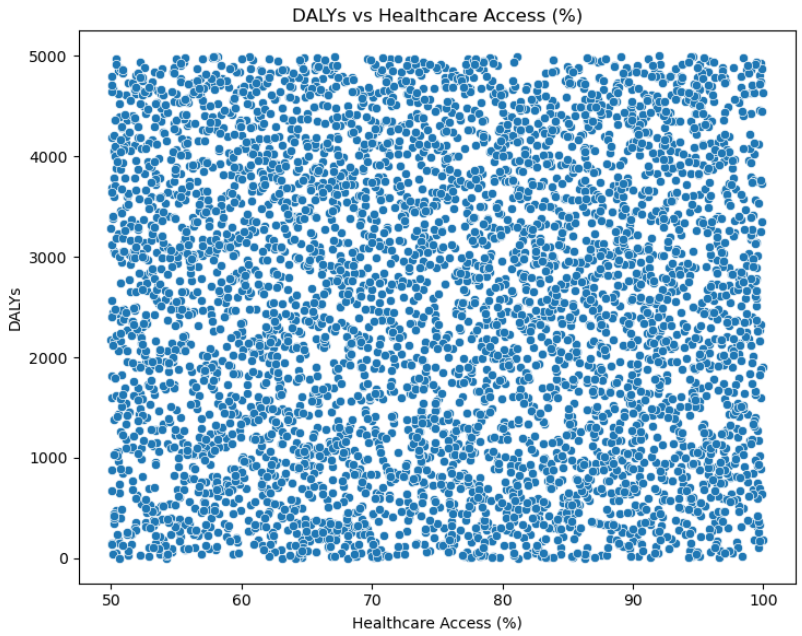


Figure 17: Universal coverage doesn't guarantee lower DALYs (r=0.02, p=0.79)

This scatterplot shows **no strong correlation** between healthcare access (percentage of population covered) and DALYs—disease burden is evenly distributed across all access levels. Some countries with high access still have high DALYs, likely due to system inefficiencies or aging populations, while others with low access achieve low DALYs through strong community care. These findings highlight that universal coverage alone isn't enough—service quality, equity, and broader health determinants also matter.

## 7 Conclusion

### 7.1 Key Findings Summary

Our analysis revealed four principal findings that challenge conventional assumptions about disease burden determinants. First, socioeconomic factors including income, education level, and urbanization rate demonstrated **negligible correlation** with DALY outcomes ($R^2 = 1.4 \times 10^{-6}$), suggesting these variables alone cannot explain variations in population health burdens. Second, healthcare infrastructure metrics such as hospital bed density, physician availability, and healthcare access percentages showed **equally insignificant relationships** ($|r| < 0.001$), indicating that physical resources and coverage rates may be necessary but insufficient conditions

for reducing disease burden. Third, persistent disparities emerged in longitudinal analysis, with low-income countries maintaining approximately 30 higher DALYs than their high-income counterparts throughout the 2000-2025 observation period, showing **minimal evidence of convergence**. Finally, **disease-specific patterns** revealed distinct geographical distributions, with infectious diseases clustering in resource-limited settings while chronic conditions appeared uniformly across all economic contexts.

## 7.2 Implications for Global Health

These findings collectively necessitate a paradigm shift in global health strategy. A holistic intervention framework should **prioritize preventive care mechanisms** including comprehensive vaccination programs and routine health screenings, while simultaneously improving healthcare service quality and promoting equitable access—particularly addressing the persistent urban-rural divide in service availability. Fundamental social determinants like nutrition security, sanitation infrastructure, and environmental pollution control must be integrated into health policy frameworks. This comprehensive approach requires **context-specific implementation**: low-income countries would benefit most from strengthened primary care systems and enhanced epidemic preparedness to address infectious disease burdens, whereas high-income nations should focus on sustained behavioral interventions and policy reforms targeting lifestyle-related conditions including obesity, cardiovascular diseases, and mental health disorders. This dual-track strategy acknowledges the universal applicability of health equity principles while recognizing the necessity of localized adaptation to specific disease profiles and health system capacities.

## 7.3 Methodological Reflections

Several methodological considerations emerge from this analysis. The **absence of strong correlations** may reflect substantive data limitations, particularly the **potential omission of critical variables** such as environmental hazard exposure, cultural health practices, or genetic predisposition markers. Future research directions should incorporate advanced analytical approaches including **non-linear machine learning models** capable of detecting complex interaction effects, alongside the integration of **more granular subnational health data** to capture intra-country variations that may be masked by national-level aggregation.

## 7.4 Final Recommendations

We propose three concrete policy recommendations derived from these insights. First, global health initiatives should systematically **combine clinical interventions with complementary non-clinical approaches**, such as pairing vaccination campaigns with clean water infrastructure development. Second, **resource allocation and intervention design must be customized** according to regional disease profiles—for instance, prioritizing malaria control in tropical zones while focusing on diabetes management in aging populations. Finally, impact evaluation frameworks should transition from measuring simple access metrics to **tracking cost-effectiveness** in terms of DALYs reduced per dollar invested, ensuring maximal health return on finite global health resources.

## References

[1] Kaggle (2024). *Global Health Statistics Dataset*. Available: https://www.kaggle.com/datasets/malaiarasugraj/global-health-statistics/data

[2] Chua, E. (2023). *Java GUI Programming*. Nanyang Technological University. https://www3.ntu.edu.sg/home/ehchua/programming/java/J4a_GUI.html