MGMT 59000-144: Computing For Analytics

Final Project Report

Topic: Data-Driven Strategies for Maximizing Global Life Expectancy

Team:
Ajay Shankar (shanka65@purdue.edu)
Dilshad Shaik (dshaik@purdue.edu)
Krupasankari Ragunathan (kragunat@purdue.edu)
Mokshda Sharma (sharm879@purdue.edu)
Rohan Ajay (rajay@purdue.edu)
Rohan Raveendran (rraveen@purdue.edu)
Saquib Hussain (hussai71@purdue.edu)
Tanvir Ahmed Farook (tfarook@purdue.edu)

Abstract
This report synthesizes key findings from a multi-technique analysis of a global health and economic dataset (179 countries, 2000-2015). Utilizing linear programming, graph algorithms, dynamic programming, K-Means clustering, robust regression, and sorting/searching, we identified critical patterns in health investment, improvement trajectories, national profiling, and wealth-health disparities. The aim is to provide actionable, data-driven intelligence for understanding global well-being.

Dataset Description

The primary dataset utilized for this analysis is the **"Life Expectancy (WHO Updated)"** dataset, publicly available on Kaggle. This dataset encompasses a range of health, economic, and demographic indicators for **179 countries**, with data spanning the years **2000 to 2015**, totalling **2,864 individual records**. The central outcome variable under investigation is 'Life expectancy'. This is supported by a rich array of features, including, but not limited to:

- **Health Indicators:** Infant mortality, adult mortality, under-five deaths, BMI, incidence of HIV, and immunization coverage rates (e.g., Hepatitis B, Polio, Diphtheria, Measles).
- **Socio Economic & Demographic Indicators:** GDP per capita, population size, average years of schooling, alcohol consumption, and economic status (developed/developing).

To augment the economic dimension, this dataset was combined with **Gross Domestic Product**

(GDP) data from the World Bank (Indicator: NY.GDP.MKTP.CD), providing a more comprehensive view of national economic capacities.

This combined dataset forms the foundation for exploring critical questions related to global health patterns, the efficiency of health improvements, optimal resource allocation, the influence of various factors on life expectancy, and regional health disparities.

Datasets

- Life Expectancy Data (Kaggle)
- GDP World Bank Data

Below is a summary of some key columns in the combined dataset:

Column Name	Description
Country	Name of the country
Region	Geographical region of the country
Year	Year of observation (2000-2015)
Life_expectancy	Life expectancy at birth (in years) - the primary outcome variable
Infant_deaths	Number of infant deaths per 1000 live births
Adult_mortality	Adult mortality rate per 1000 population (probability of dying between 15 and 60 years)
Under_five_deaths	Number of under-five deaths per 1000 population
BMI	Average Body Mass Index of the population
Schooling	Average number of years of schooling
GDP_per_capita	Gross Domestic Product per capita (current US$)
HIV_AIDS	Deaths per 1000 live births due to HIV/AIDS
Measles	Number of reported measles cases per 1000 population
Hepatitis_B	Percentage of 1-year-olds immunized against Hepatitis B
Polio	Percentage of 1-year-olds immunized against Polio
Diphtheria	Percentage of 1-year-olds immunized against Diphtheria
Population	Population of the country
Status	Developed or Developing status of the country
Alcohol_consumption	Alcohol consumption in litres of pure alcohol per capita (≥15 years old)

This comprehensive dataset allows for a multi-faceted exploration of factors influencing global health. Leveraging this rich, multi-dimensional dataset, we explored a range of analytical and optimization problems aimed at understanding global health patterns, the efficiency of health improvements, optimal resource allocation, and regional health disparities. Each problem targets a different aspect of public health analysis from optimizing budget allocations using linear programming, to identifying periods of maximum health improvement using dynamic programming, and uncovering country groupings and disparities through clustering and search algorithms. The following sections outline these problem statements, the techniques applied, and the insights generated from each approach.

Objective #1: Optimizing Health Budget Allocation for Maximum Life Expectancy Gain

To determine the most effective way to allocate a limited health budget across different public health interventions to maximize life expectancy. Specifically, the study explored how investments in reducing infant mortality, improving nutrition (BMI), and enhancing access to education (schooling) influenced average life expectancy.

Methodology: Linear Programming (LP)

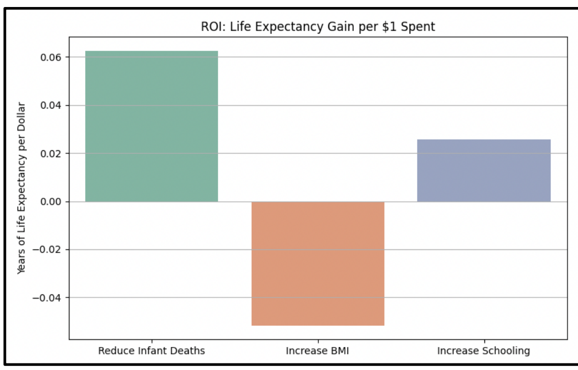
The analysis followed a structured, three-phase process:

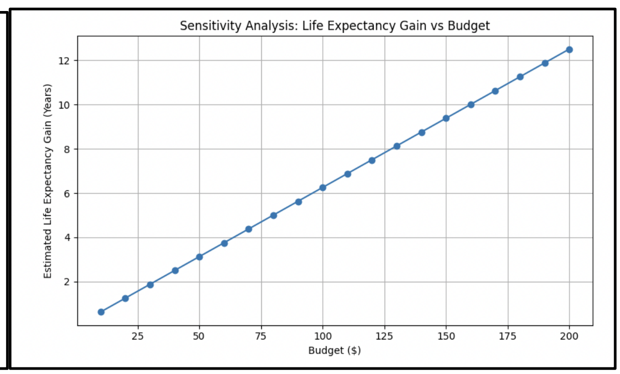
- **Phase 1: Data Preparation and Regression Modeling**
  - A life expectancy dataset was cleaned, focusing on non-missing values for Infant_deaths, BMI, Schooling, and Life_expectancy.
  - A linear regression model was trained to predict Life_expectancy using the other three variables. The resulting regression coefficients quantified the marginal effect of each:
    - **Infant_deaths:** -0.3127
    - **BMI**: -0.1035
    - **Schooling**: +0.0766
- **Phase 2: Linear Programming (LP) Optimization**
  - The regression coefficients were used in an LP model (using PuLP library) to maximize life expectancy gain under a strict budget of $100.
  - Decision variables represented units of improvement for each factor, with associated costs: $5 for reducing infant deaths, $2 for improving BMI, and $3 for increasing schooling.
  - The objective was to maximize the weighted sum of life expectancy gains, constrained by the total budget.
- **Phase 3: Insights and Visualization**
  - Return on Investment (ROI) was calculated (years of life expectancy gained per dollar).
  - A sensitivity analysis was performed to see how total expected life expectancy gain

changes with increasing budget levels.

Objective #1 Insights

- **Optimal Allocation:** The LP solution identified that 100% of the available budget should be allocated to **reducing infant deaths**, resulting in a 6.25-year increase in life expectancy for a $100 budget.
- **Highest ROI:** Reducing infant deaths offers the highest return per dollar, yielding approximately **0.0625 years per dollar invested**. In contrast, schooling offers a smaller positive return (~0.0255 years per dollar), and BMI interventions showed a negative return on investment in this context.
- **Scalability:** The sensitivity analysis demonstrated a linear and scalable benefit from increasing budget levels when investing exclusively in reducing infant mortality (e.g., doubling the budget to $200 would result in a 12.5-year gain).
- **Confirmation:** The regression model confirmed that infant mortality has the largest negative influence on life expectancy, reaffirming its importance as a policy priority.





Objective #2: Identifying Maximum Sustained Improvements in National Life Expectancy

To understand how much life expectancy countries typically gain during their most successful, uninterrupted periods of improvement, and to identify which countries have achieved the most significant long-term gains.

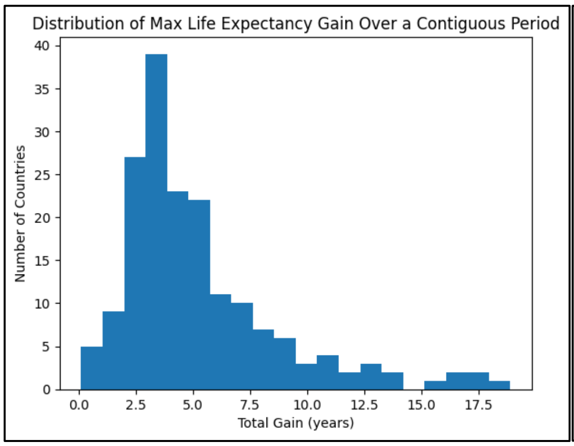
Methodology: Dynamic Programming (DP)

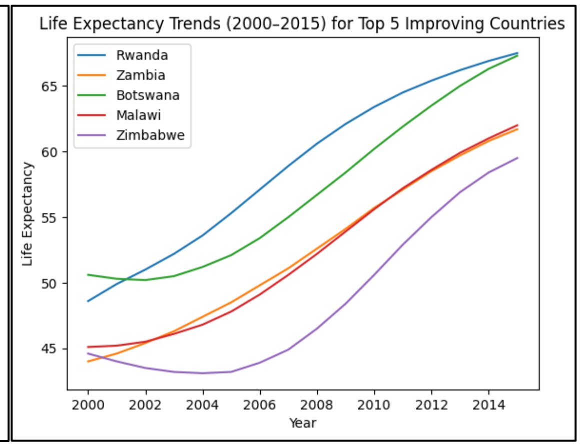
- **Step 1: Calculate Annual Changes:** For each country, the year-to-year change (delta) in life expectancy was calculated for the period 2000-2015. delta[t] = LifeExpectancy[t+1] – LifeExpectancy[t]
- **Step 2: Apply Dynamic Programming:** A DP algorithm was used to find the contiguous span of years for each country where the sum of these annual deltas (gains in life expectancy) was maximal. This represents the country's "best run" of improvement. dp[i]

= max(delta[i], dp[i-1] + delta[i])

- **Step 3: Analyze Distributions and Trends:**
  - A histogram of these maximum gains across all countries was plotted to show the distribution of improvements.
  - The top 5 countries with the highest DP-found gains were identified and their life expectancy trends plotted.

Objective #2 Insights





- **Typical Gains are Modest:** For most countries, their best period of uninterrupted improvement resulted in a life expectancy gain of around **2-6 years**.
- **Outlier Gains for Low-Starting Countries:** A few countries, often starting with very low life expectancy, achieved much larger gains of **10-18 years** during their best continuous improvement streak. This suggests they had more "low-hanging fruit" to address.
- **Top Performers:** Countries like Rwanda (approx. +19 years from 2000 to 2015), Zambia, Botswana, Malawi, and Zimbabwe showed very steep climbs. This often followed periods of major health crises, with subsequent sustained interventions (e.g., child immunizations, HIV treatment scale-up) enabling long streaks of positive change.
- **Pattern:** Lower starting life expectancy correlated with higher possible DP-found gains. Countries already in the 70s or 80s for life expectancy tended to inch forward, with their best DP segments netting only a few years.

Objective #3: Investigating the Relationship Between Economic Wealth and Public Health

To determine whether a nation's economic prosperity (measured by GDP per capita) directly and consistently translates into better public health outcomes (measured by life expectancy).

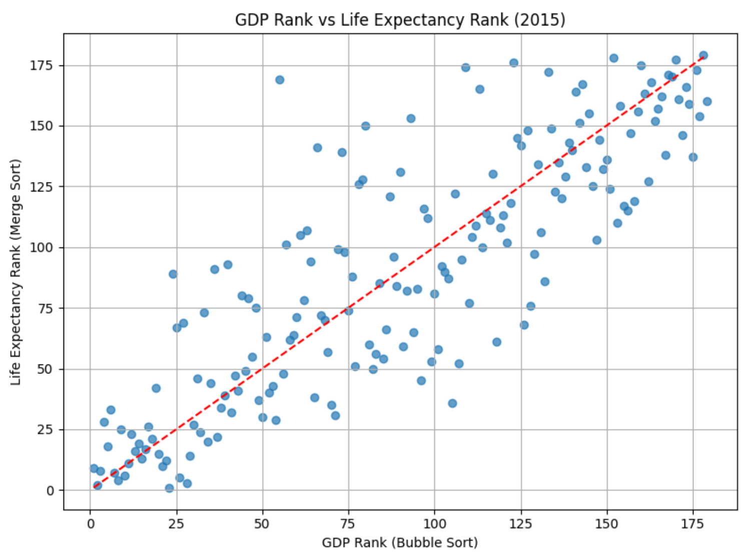
Methodology: Sorting and Comparative Ranking

- **Step 1: Data Compilation:** Data on GDP per capita and Life Expectancy for various countries were compiled.

- **Step 2: Ranking:** Countries were ranked independently based on their GDP per capita and their life expectancy.
- **Step 3: Discrepancy Analysis:** The ranks were compared to identify countries where there was a significant mismatch between economic wealth ranking and health outcome ranking. This was visualized (as mentioned in the source text, likely via a discrepancy chart).

Objective #3 Insights

- **Wealth is Not a Guarantee of Health:** The analysis clearly showed that economic wealth doesn't always translate directly to better public health.
  - Some countries with very **high GDP per capita** still had relatively **low life expectancy**.
  - Conversely, some countries with **modest GDP** achieved **excellent health outcomes** and high life expectancy.
- **Implication:** This suggests that while economic resources are a factor, other elements such as the effectiveness of the healthcare system, public health policies, equitable distribution of resources, education, and social determinants of health play critical roles in achieving high life expectancy.

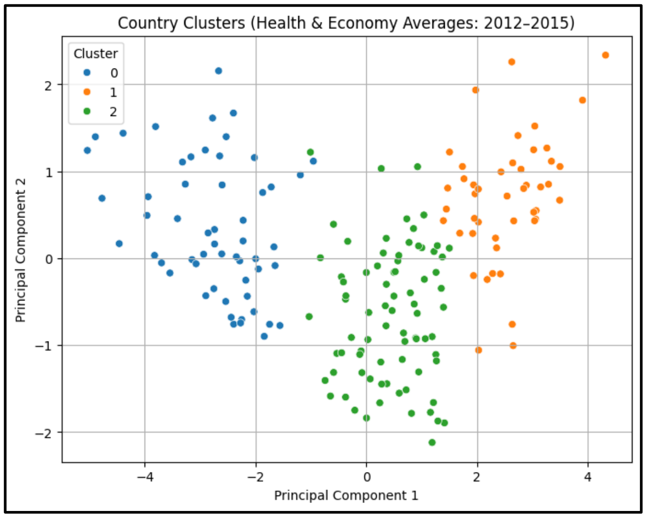


Objective #4: Discovering natural groupings of countries without predefined labels

To ascertain if countries can be naturally grouped into distinct clusters based on their average health and economic indicators, revealing broader global patterns of development, risk, and health status.

Methodology: K-Means Clustering

- **Step 1: Data Aggregation and Selection:** Health and economic indicators (including GDP per capita, life expectancy, schooling levels, infant and adult mortality rates) were averaged for each country over the period 2012-2015.
- **Step 2: Feature Scaling:** Selected features were standardized (e.g., using StandardScaler) to ensure that variables with larger magnitudes did not disproportionately influence the clustering process.
- **Step 3: K-Means Algorithm Application:** The K-Means clustering algorithm was applied to the scaled data. An appropriate number of clusters (k=3 was identified in the analysis) was determined, potentially using methods like the elbow method or silhouette analysis.
- **Step 4: Cluster Interpretation:** The characteristics of the countries within each cluster were analyzed to understand the common traits that defined each group.



Objective #4 Insights

Using K-Means clustering, countries were naturally grouped into three distinct clusters based on their average health and economic indicators between 2012 and 2015. These clusters revealed clear and interpretable patterns:

- **Cluster 1 (Wealthy and Healthy):** Included countries with high GDP per capita, high life expectancy, strong education levels, and low mortality rates. These are typically developed nations with robust health systems.
- **Cluster 2 (Mid-Tier Nations):** Grouped countries with moderate GDP and mixed health outcomes, often reflecting emerging economies in transition. These nations may show improving but still variable health indicators.
- **Cluster 0 (Low-Income, High-Risk):** Captured low-income countries characterized by poor health indicators, low schooling attainment, and high infant and adult mortality rates. These nations face significant public health challenges

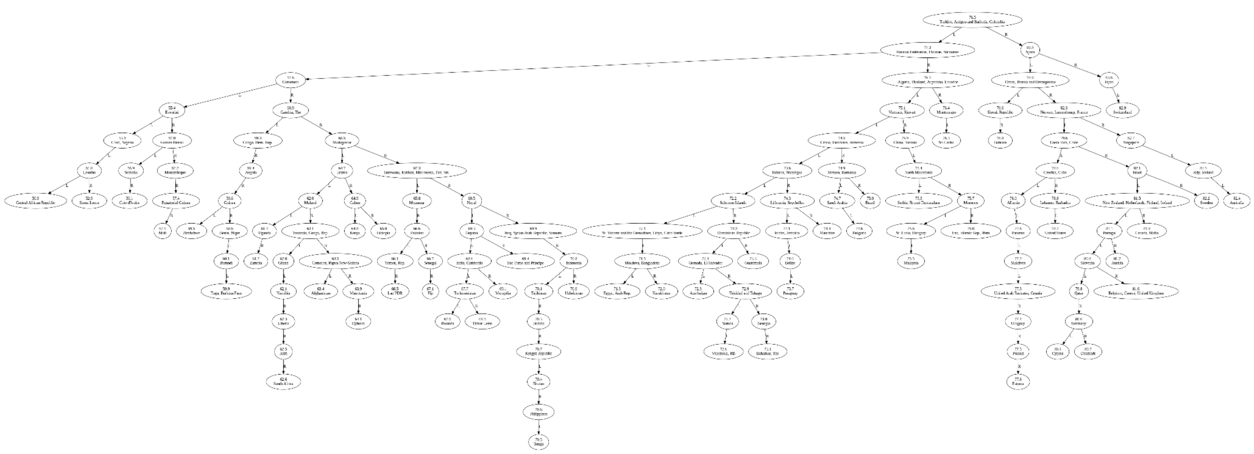
Objective #5: Uncovering Socio-Economic Patterns within Life Expectancy Tiers

To explore whether countries falling into specific life expectancy brackets exhibit common socio-economic characteristics or face similar developmental challenges and opportunities.

Methodology: Searching (using Binary Search Tree principles for grouping)

- **Step 1: Data Segmentation:** Countries were implicitly or explicitly grouped based on their life expectancy values, creating tiers or ranges (e.g., life expectancy below 60, mid-range life expectancy). While a Binary Search Tree (BST) was mentioned as the conceptual tool, this likely involved sorting by life expectancy and then analyzing segments.
- **Step 2: Profile Analysis within Segments:** For each segment/tier, associated data on GDP, healthcare system development, education investments, and urbanization were examined to identify common patterns or trends.
- **Step 3: Comparative Insights:** The socio-economic profiles of different tiers were compared to draw insights about the varying conditions associated with different levels of life expectancy.

Objective #5 Insights



Using grouping based on life expectancy ranges (conceptually aided by Binary Search Tree structures for finding natural groupings), distinct patterns emerged:

- **Mid-Range Life Expectancy Countries:** Nations like Ethiopia, Myanmar, and Lao PDR, often falling into a mid-range for life expectancy, frequently showed signs of **improving healthcare systems, growing investments in education, and lower levels of urbanization** compared to highly developed nations. This suggests a transitional phase of development.
- **Countries with Life Expectancy Below 60:** For countries grouped with life expectancy below 60 years, a critical pattern observed was that **increases in GDP do not consistently or effectively trickle down to tangible improvements in healthcare systems or social services.** This points to potential systemic issues in resource allocation, governance, or

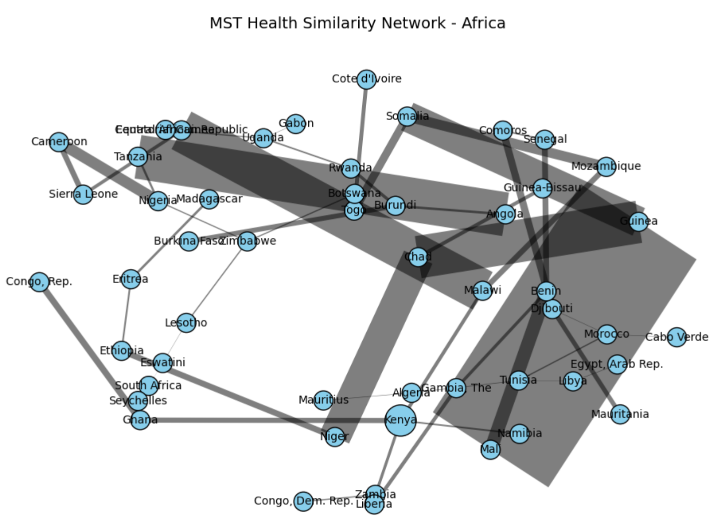
infrastructure that hinder the conversion of economic growth into public health benefits for the most vulnerable populations.

Objective #6: Identifying Structurally Central Countries in Regional Health Networks

To identify countries within specific geographical regions that, while not necessarily the most developed or highest-performing, are structurally central in terms of their health and development profiles relative to their neighbors. These "exemplar" countries can be key for understanding regional trends and coordinating efforts.

Methodology: Graph Algorithms (Minimum Spanning Tree - MST)

- **Step 1: Data Preparation and Regional Subsetting:** The dataset was filtered for specific geographic regions (e.g., Oceania, Rest of Europe, South America). Relevant health and development indicators were selected for similarity comparison.
- **Step 2: Graph Construction:**
  - Countries within a region were treated as nodes in a graph.
  - A similarity metric (e.g., Euclidean distance based on standardized health/development indicators) was calculated between all pairs of countries within the region. Edge weights were typically inverse to similarity (lower weight for more similar countries).
- **Step 3: Minimum Spanning Tree (MST) Construction:** For each region, an MST was constructed. The MST connects all countries in the region with the minimum possible total edge weight, highlighting the strongest relationships or similarities.
- **Step 4: Centrality Analysis:** Countries with a high degree (number of connections) in the MST were identified as "exemplar" or structurally central countries. These are nodes that connect multiple other nodes with strong similarity links.



Objective #6 Insights

- **Structural Centrality Revealed:** MST analysis successfully identified countries that are structurally central within their regional health networks. These countries are not necessarily the most advanced but represent key "hubs" or typical profiles for their region.
- **Exemplar Countries for Benchmarking and Collaboration:**
    - **Oceania:** Tonga was identified as an exemplar country (degree of 3 in MST), suggesting its health profile is central to the Oceanic region and may reflect common regional patterns.
    - **Rest of Europe (non-EU):** Moldova (degree of 3 in MST) emerged as a representative hub, indicative of shared public health and socioeconomic contexts among non-EU European nations.
    - **South America:** Chile (degree of 3 in MST) showed centrality, suggesting it shares strong similarities in key health and development dimensions with its neighbors.
- **Applications:** These exemplar countries are ideal candidates for regional benchmarking, policy comparison, and planning collaborative public health initiatives, as they can provide insights into common challenges and effective strategies within their specific regional contexts.

Summary and Conclusion

1. **Optimal Investment Targets:**
   Linear programming showed that, under a fixed budget, allocating resources exclusively to reducing infant mortality maximizes life-expectancy gain, delivering roughly 0.0625 years per dollar spent which far outpaces investments in BMI or schooling.
2. **Patterns of Improvement Over Time:**
   Dynamic programming revealed that most countries' best uninterrupted life-expectancy streaks yield 2–6 years of gain, whereas low-starting nations can achieve 10–18 years during their fastest improvement phase.
3. **Wealth vs. Health Disparities:**
   Sorting and rank-discrepancy analysis confirmed that higher GDP per capita does not guarantee better health outcomes. Some affluent nations lag in life expectancy while several lower-income countries outperform their economic peers.
4. **Natural Country Groupings:**
   K-Means clustering uncovered three distinct clusters- *"Wealthy & Healthy," "Mid-Tier,"* and *"Low-Income High-Risk",* highlighting broad development and health-status patterns across regions.
5. **Life-Expectancy Tier Profiles:**
   Segmenting countries by life-expectancy brackets showed that mid-range nations often enjoy improving healthcare and education infrastructure, whereas countries below 60 years struggle to translate GDP growth into tangible health gains, pointing to governance and allocation

challenges.

6. **Regional Health Hubs:**
   Minimum spanning-tree (MST) analysis identified "exemplar" countries, such as Tonga in Oceania, Moldova in non-EU Europe, and Chile in South America that serve as structural centers in regional health networks and can inform benchmarking and collaborative policies.

A more effective way to improve global life expectancy is to focus on a few key areas: reducing infant deaths, supporting countries that are steadily improving, and understanding that having a strong economy doesn't always mean better health. Governments should invest in high-impact solutions like child vaccination and maternal care, help middle-performing countries improve health and education, and learn from countries that serve as regional examples. By targeting these areas, we can reduce health gaps and help people live longer lives around the world.