CPSC 368 Group 10 Research Paper

Introduction

Topic Background:

In recent decades, digital technologies have significantly reshaped public health systems and healthcare delivery around the world. The widespread use of the internet has enabled improved access to health information, telemedicine services, remote diagnostics, and more effective public health communication. However, the global distribution of internet access remains uneven, raising the question of whether digital connectivity contributes to measurable differences in healthcare outcomes across countries.

This project investigates the association between digital connectivity and healthcare quality across countries in the year 2015. Specifically, we examine whether a country's internet usage rate is correlated with key indicators of healthcare quality: life expectancy, crude death rate, and immunization rate for DPT (Diphtheria, Pertussis, and Tetanus). While the internet may not directly influence these health outcomes, greater digital access may serve as a proxy for improved access to healthcare resources, information, and infrastructure.

To ensure that observed relationships are not confounded by other factors, we also include control variables such as GDP per capita, health expenditure per capita, and urban population percentage. These factors are known to influence healthcare quality and may co-vary with digital infrastructure.

Research Question:

What is the association between digital connectivity (measured by a population's access to the Internet) and healthcare quality across countries in 2015?

Healthcare quality will be explored in three ways:

1. Diphtheria, Pertussis, and Tetanus (DPT) Immunization
    a. DPT Immunization is widely recognized as a key indicator of a country's healthcare quality, as seen in how The World Health Organization (WHO) monitors DPT3 immunization coverage worldwide [11]
2. Life expectancy
    a. We believe life expectancy to be "a core indicator in assessing public health", as argued by Aimbetova et al. [2022]
3. Crude death rate

a. The crude death rate is widely recognised as “a good indicator of the general health status” of a country, as seen by the World Bank Group [1]

This research question is important because it explores how digital connectivity may contribute to better health outcomes globally. By examining three distinct indicators - preventive care (DPT immunization), overall population health (life expectancy), and mortality (crude death rate) - the study can address **multiple aspects** of healthcare quality.

Impact:

The results of this research can be used to identify the role of digital connectivity in improving healthcare in countries. Policymakers can utilize this information to make informed decisions and better allocate their budget. Additionally, if increased digital connectivity does lead to improved health quality for its citizens, we can examine **why** that is the case and use this information to better the healthcare of a country. For example, if increased digital connectivity leads to better health care through telemedicine services, the country can allocate more resources in that area.

Related work

The relationship between internet access and healthcare outcomes has been examined across multiple academic disciplines, including health economics, public health, and digital policy. Existing research broadly supports the idea that digital connectivity can improve access to care, enhance service efficiency, and help reduce health disparities—though outcomes often depend on wider social and infrastructural contexts.

Van Parys and Brown (2023) studied the effects of broadband internet access within the U.S [4]. Medicare system, focusing on both patients and healthcare providers. Their analysis found that increased broadband availability was associated with improved healthcare utilization and patient outcomes, particularly in rural and underserved areas. These findings support our study’s hypothesis that internet access can serve as a proxy for broader access to health information and services.

Yu and Meng (2022) conducted a cross-country analysis to explore how internet access influences healthcare access and inequality [5]. Their results suggest that greater internet penetration tends to reduce disparities in healthcare access—especially when combined with public health initiatives. Our study builds on this global perspective by using standardized 2015 data from the World Bank and United Nations to examine similar associations across national contexts.

Earlier work by Kerwin and James (2002) investigated the qualitative and organizational effects of internet technologies on healthcare quality [6]. They identified improvements in care coordination, electronic recordkeeping, and provider-patient communication as early benefits of digital adoption. While their study was conducted in a different technological era, it offers valuable foundational insights into the mechanisms through which digital access may influence national health indicators such as life expectancy and immunization rates.

Collectively, these studies demonstrate that internet access has the potential to positively influence healthcare delivery and outcomes. By incorporating multiple global health indicators and controlling for key socioeconomic variables, our project contributes a broad, data-driven perspective to this growing body of research—grounded in publicly available country-level data from 2015.

Data

Dataset Overview:

The first data set is sourced from The Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat. Contains world mortality rates and life expectancy data by country in the 2015 revision of the World Population Prospects. [10]

url:
https://digitallibrary.un.org/record/831127/files/World_Mortality_2015_wall_chart_FINAL.pdf

The second data set comes from the World Bank Group, an international organization which supports developing countries through financial loans. This database contains global monetary indicators from over the last 50 years. [9]

url: https://datatopics.worldbank.org/world-development-indicators/

The UN data set on world mortality is sourced majorly from census data and surveys from each country/region. Issues such as underreporting and inconsistent government statistics can affect the accuracy of the dataset. In developing regions where data may be less reliable, estimates were made using neighbouring regions' data with equivalent socio-economic status, at times in conjunction with local surveys and population registers.

However, despite these challenges, the dataset remains a valuable and trustworthy resource due to its rigorous methodology and global standardization. The combination of multiple sources, cross-validation with international data sets, and the use of probabilistic methods enhance the reliability of the projections.

The World Bank's World Development Indicators (WDI) dataset is compiled from officially recognized international sources, including national statistical agencies, central banks, and international organizations. While this comprehensive approach ensures extensive coverage, it may introduce biases due to variations in data collection methodologies, inconsistent reporting standards, and potential political influences across countries. In cases where data is missing or unreliable, the World Bank employs estimation techniques, often using regional trends or related economic indicators to fill gaps.

However, the World Bank remains a highly credible and trustworthy source due to its strict validation processes and reliance on internationally recognized methodologies. The organization collaborates with global agencies such as the United Nations and national governments to ensure data consistency and comparability. Its commitment to transparency allows researchers to access metadata, methodologies, and confidence intervals, enabling independent verification.

We choose to cross-reference these two datasets by identifying overlapping country-level data and analyzing discrepancies. The infant mortality and maternal mortality rates in both datasets generally align, but some variations exist due to differences in data collection methods and estimation models. The UN dataset primarily relies on census data, household surveys, and national health records, while the WBG often incorporates modelled estimates and interpolations to compensate for missing or unreliable reports.

The acquisition methodology for the data can be found here:

UN : https://population.un.org/wpp/assets/Files/WPP2015_Methodology.pdf
WBG: https://datatopics.worldbank.org/world-development-indicators/sources-and-methods.html

Data Cleaning:

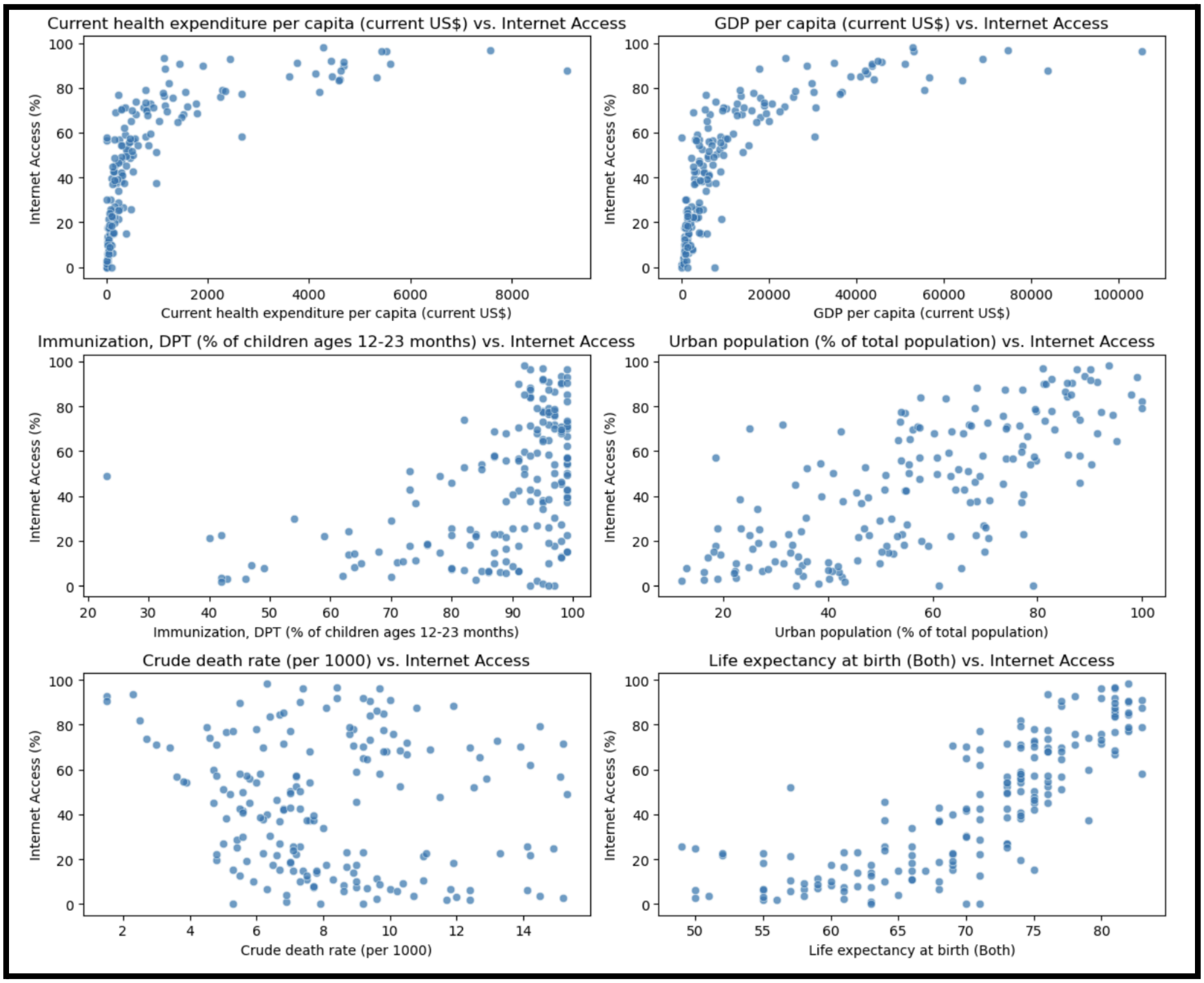
World Bank Data: Since the UN data is from 2015, we first isolated the World Bank dataset to the year 2015. The data is in a long format, where each row contains an indicator and the rows are grouped by country or region. Since we were only looking at data from one year, we pivoted the table so that each row is a unique country, and each column is a single indicator. Each cell contains a single value, and the dataset was then in a tidy format. From there, we isolated our indicators of interest: DPT Immunization rate and % of peoples using the internet. Additionally, we kept indicators to evaluate for confounding variables: current health expenditure per capita, GDP per capita, and urban population %.

UN Data: We started by converting the data from the pdf file into an Excel spreadsheet. Since the data was already in a tidy format, there was no pivoting needed.

We then isolated the country data in both datasets by removing the region or continent data. We decided to use the 195 countries recognized by the United Nations (193 member states and 2 observer states) as a standard, and therefore removed all territories and constituencies. In order to merge the data we verified that each dataset had the same countries. There were several that were missing from the UN dataset (for reasons unknown): Andorra, Dominica, Kosovo, Liechtenstein, Marshall Islands, Monaco, Nauru, San Marino, St. Kitts and Nevis, and Tuvalu. These were therefore removed from the World Bank Data.

Exploratory Data Analysis (EDA):

The visualisation below is the scatter-plot of the explanatory variables and the controlled confounding variables, versus our target variables (Internet Access). It is meant to examine if there is any potential relationship between Internet Assess and the potential indicators.



Given the results of the EDA, these are the following steps:

1. Healthcare Expenditure and GDP per capita:
    a. Internet access seems to increase with the following variables, but the trend is non-linear

    b. Therefore, log transformation will be performed on these variables in order to make the trend linear. This is necessary for our intended method, multiple linear regression, as it improves the model's accuracy
2. Urban Population and Life Expectancy:
    a. Internet access seems to increase with the following variables, and the trend is linear
    b. Therefore, we could leave these variables as is
3. DPT Immunization:
    a. All the points seem to be clustered at the higher values, not allows us to observe any trend
    b. Therefore, we can further analyze the distribution of this variable by making a histogram. And if the distribution seems very skewed, we can standardise the variable, or apply min-max scaling on it
4. Crude Death Rate:
    a. Internet access seems to decrease as crude death rate increases, and the trend is linear. But, the correlation seems to be very weak
    b. Therefore, we could explore interaction terms instead (eg. Crude death rate * GDP per capita)

Methodology

1. **Exploratory Data Analysis (EDA):** As done above, we checked our data to see the distribution of the data, and let us identify any potential trends, outliers, or missing data. And by doing so, we will now perform the necessary transformations on our data.

2. **Confounding Variables:** To ensure that the analysis focuses on the relationship between digital connectivity and healthcare, we will control possible confounding variables. The confounding variables we will control include GDP, healthcare expenditure, and urbanization rates.

3. **Regression:** We will use multiple linear regression to see the association between digital connectivity and each healthcare quality indicator. Therefore, our regression model will look like this:

$$Y = \beta_0 + \beta_1(Internet\ Access) + \beta_2(GDP\ per\ capita) + \beta_3(Healthcare\ Expenditure) + \beta_4(Urbanization\ Rate) + \epsilon$$

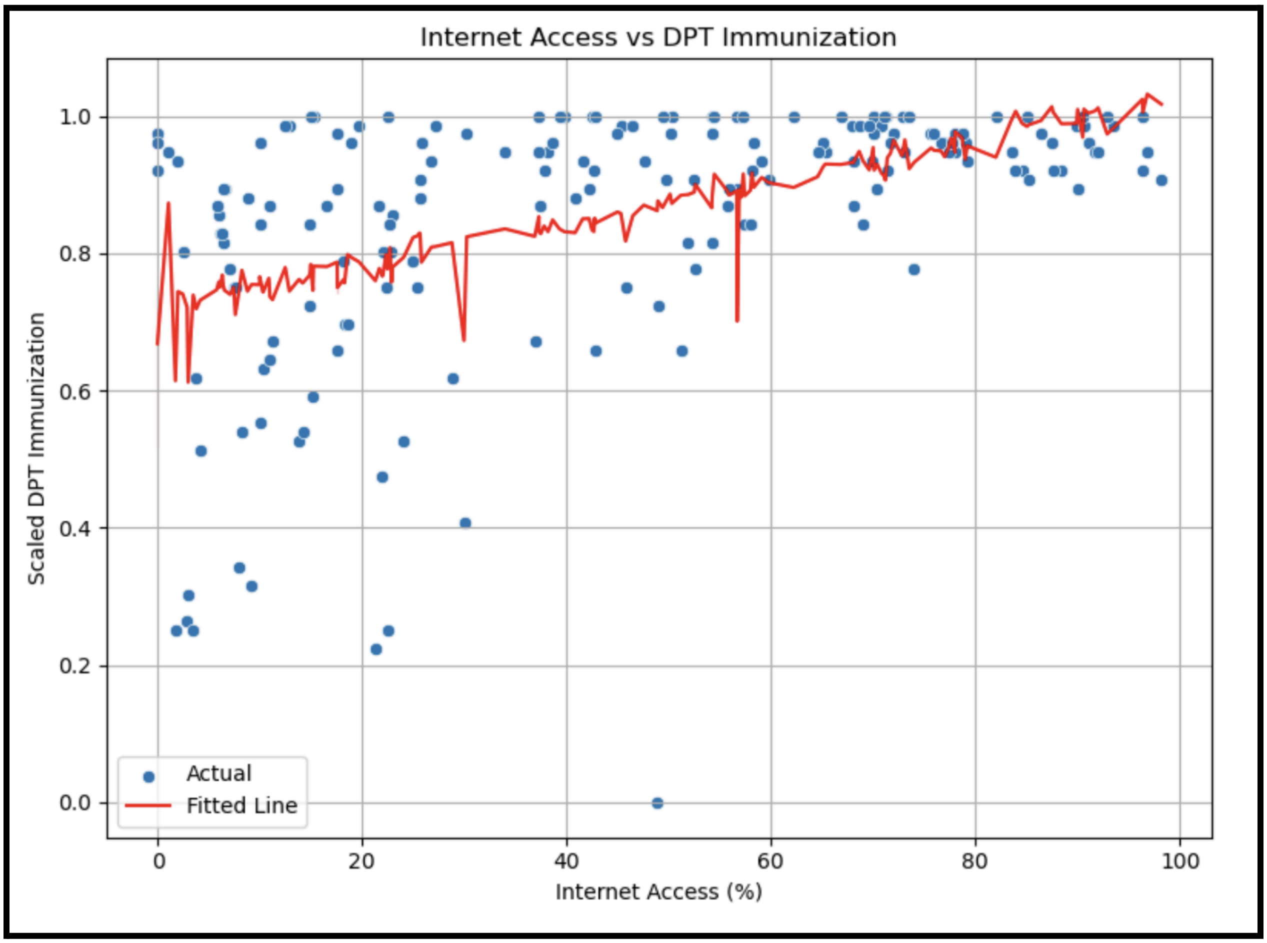
   where Y is the health indicator. For any healthcare indicator, if the coefficient for Internet Access, β_1, is statistically significant, it can suggest a meaningful association between digital connectivity and healthcare quality.

4. **Assess Model Performance:** We will check the adjusted R^2, or measure the Mean Square Error to assess the model's performance. Additionally, we will also check the p-value (if it's below 0.05) for statistical significance.
5. **Ensure soundness:** We can ensure that the method is sound by performing the necessary transformations on our variables after EDA (such as log transformation or min-max scaling).

The pros of this methodology is that possible confounders are controlled, and that statistical significance is rigorously checked. The cons however is that linear regression assumes linearity and normality in the data.

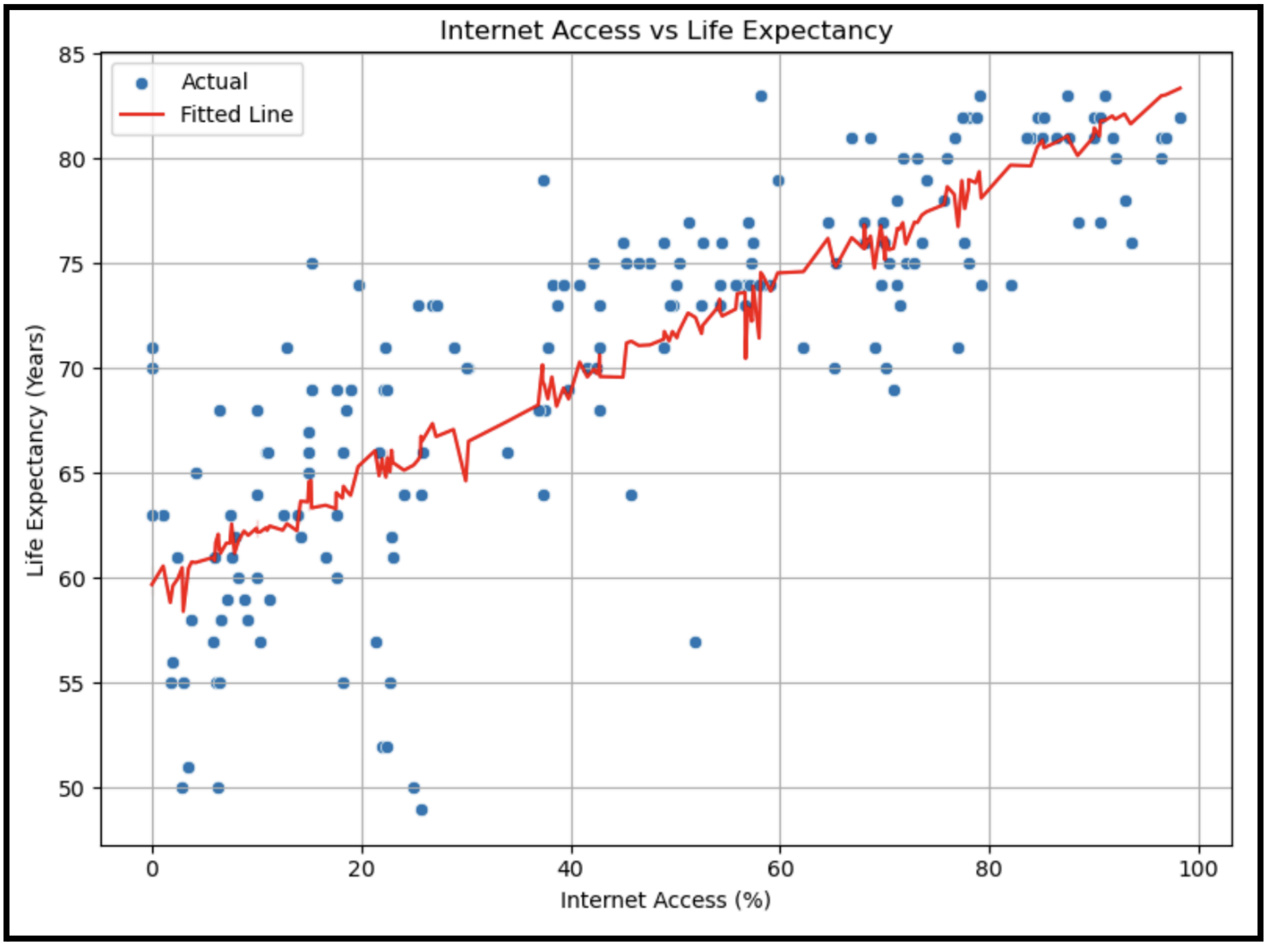
Results

To investigate the relationship between digital connectivity and healthcare quality across countries in 2015, multiple linear regression models were constructed with Internet access as a key predictor and three health-related outcomes: DPT immunization rates, life expectancy, and crude death rate.

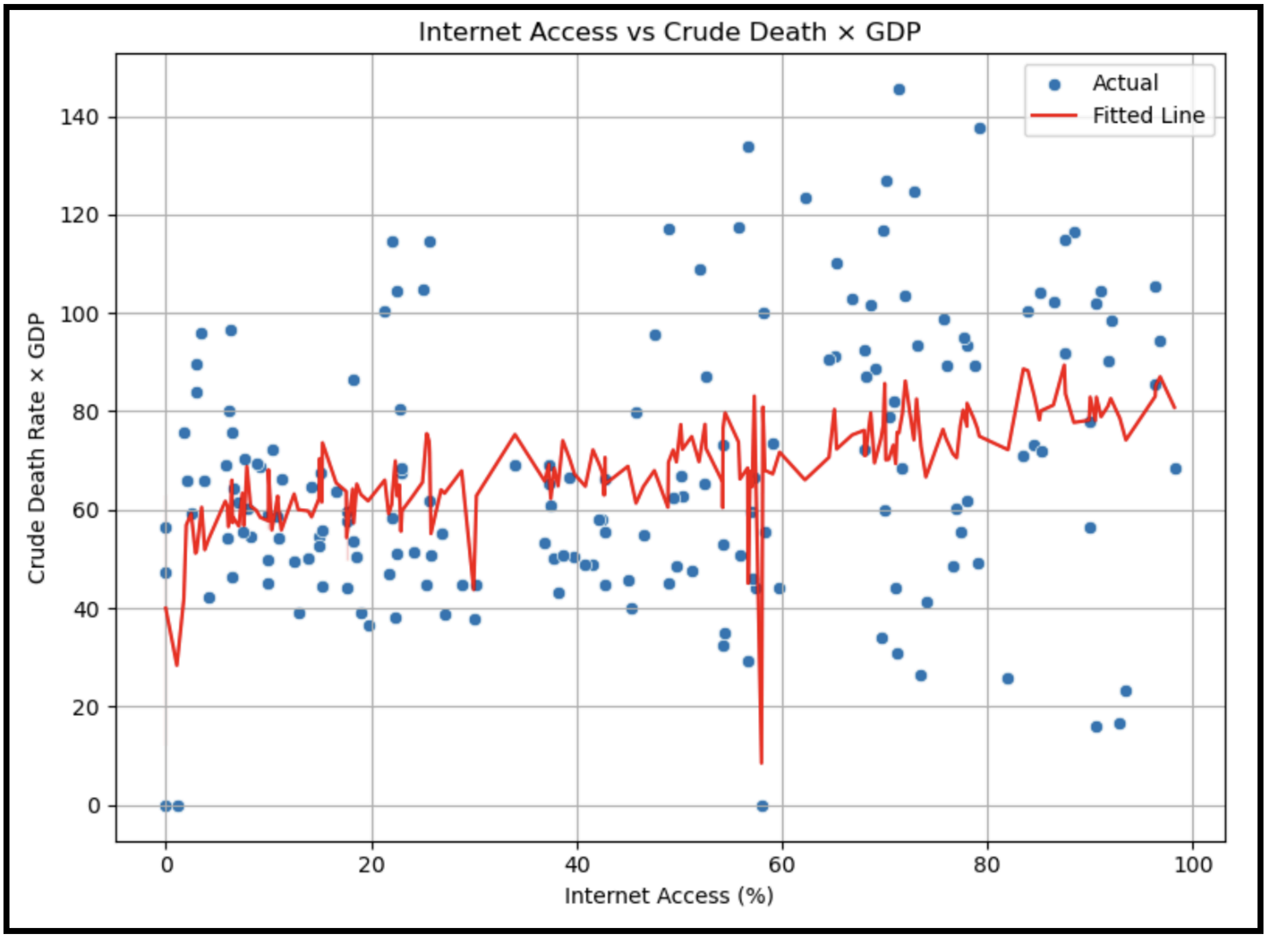


Internet access was positively and significantly associated with DPT immunization rates, even after adjusting for GDP per capita, health expenditure, and urban population. A 1% increase in

Internet access was associated with a 0.0028 unit increase in scaled immunization rates. This could suggest that digital connectivity may enhance the reach of public health campaigns. However, health expenditure per capita also had a significant positive effect, showing the need to control this confounding variable to avoid overstating the role of internet access on DPT Immunization.



Internet access was positively and significantly associated with Life Expectancy, even after adjusting for GDP per capita, health expenditure, and urban population. A 1% increase in Internet access was associated with a 0.2 year increase in life expectancy. This could suggest that digital connectivity may improve access to health-related information services, like telemedicine. In contrast, the other predictors - GDP, health expenditure, and urbanization - were not statistically significant after accounting for Internet access, suggesting that digital connectivity may be more independently associated with internet access.



Internet access was not significantly associated with Crude Death Rate, after adjusting for GDP per capita, health expenditure, and urban population. This suggests that crude death rate may be more susceptible to factors like war or epidemics, something that digital connectivity may not directly influence.

Internal Validity:

For DPT immunization, the p-value was 0.002 and the 95% confidence interval did not include zero. This provides strong evidence that the relationship is not due to random chance, increasing the validity of the observed positive association between digital connectivity and DPT immunization rates

For Life Expectancy, the p-value was less than 0.001 and the 95% confidence interval did not include zero. This provides strong evidence that the relationship is not due to random chance, increasing the validity of the observed positive association between digital connectivity and life expectancy.

For crude death rate × GDP, the p-value was 0.721 and the 95% confidence interval includes zero. But it was precisely due to this lack of statistical significance that we concluded that there was no relationship between digital connectivity and crude death rate, making our conclusion valid.

External Validity:

There are many nuances to the association between immunization rates and internet usage. Much of the research on this topic has been to analyze the effects of misinformation (ex. MMR vaccines cause autism) on vaccine skepticism. A 2022 study on countries in Sub-Saharan Africa [8] examined the effects of internet users searching vaccine information, and found that higher levels of searches were associated with overall higher vaccination coverage. The association was low, which could be due to the fact that the Measles vaccine was included, which is particularly controversial. Despite the potential effects, these results generally align with our findings of a small but significant association between internet access and DPT vaccination rates.

In a 2023 study performed on 48 countries in Sub-Saharan Africa using quantile regression [2], researchers found that higher internet access was associated with longer lifespans, and lower infant mortality. The relationship, when controlled by GDP per capita, Health Expenditure, telephone usage and CO_2 emissions was found to be linear.. This analysis was performed using data from the World Bank's Development Indicators from 2000-2020. Our findings align with these results, using the same source. Despite the limited scope of this study, the similarities indicate that our results are reproducible.

Crude death rate is an all-cause mortality indicator, which is very broad. As such, there are currently no robust studies on the association between digital connectivity and crude death rate. One could look at more specific mortality indicators and their associations with internet access in order to paint a picture. The leading cause of death globally is cardiovascular disease. In a 2022 study performed using CDC data from the United States [7], researchers found that there was an inverse relationship between internet access and age-adjusted cardiac mortality. However, it is likely that crude death rate is influenced by many complex factors and cannot be simply associated with internet access. Further study is required to validate the results of our research.

Discussion

Results:

Our analysis focuses on three metrics: DPT immunization rates, crude death rate and Life expectancy at birth. In the case of DPT immunization, we find that there is a positive association between internet access and immunization rates, as seen in the adjusted R^2 value of 0.238. This is interpreted to mean that our model explains 23.8% of the variability in DPT immunization rates. These findings align with our initial EDA, where we observed a slight positive correlation between internet access and immunization. We obtain an R-squared value of 0.255, which is only slightly higher than the adjusted value. This can indicate that the model variables don’t add any unnecessary contributions, and therefore that there is little correlation between them. While both the adjusted r-squared and the r-squared values are low, this doesn't mean that the model is irrelevant. We will look at the statistical significance to gain more information about our model.

Our model finds a coefficient of 0.0028 for Internet Access, and a p-value of 0.002, which is statistically significant at a 5% significance level. While both GDP per capita and Urban Population yielded large p-values (not significant), we find a statistically significant p-value for Health Expenditure ($p = 0.019$). This indicates that both internet access and health expenditure have an effect on immunization rates. From our adjusted r-squared value above, we must also acknowledge that there is a lot of variability not explained by our model. This means that while other factors may influence immunization rates, internet access may increase awareness and/or improve access (telemedicine) to healthcare providers in conjunction with higher spending.

Our analysis of life expectancy yields more significant findings, with an adjusted r-squared value of 0.666. Our model explains 66.66% of the variance in life expectancy at birth, which is quite high. Our r-squared value of 0.673 is similar, which could indicate a number of possibilities; the most likely being that there are little or no unnecessary indicators in our model. However, this will be further discussed in our limitations. The model finds a p-value of 0.001 for internet access, which is statistically significant ($p<0.05$). All other variables have high p-values and are therefore not statistically significant. It can therefore be said that internet access is a good indicator of life expectancy at birth, and is independent from GDP, urban population % and health expenditure. This aligns with the distribution in our exploratory data analysis, which showed a clear positive correlation between internet access and life expectancy. These results underscore how important digital connectivity is for the health of a population, despite a country's wealth and/or infrastructure.

Our final model, which examines the relationship between crude death rate and our indicators, yields weak results. We find a low adjusted r-squared value of 0.162, and an r-squared value of 0.181. Our model therefore explains a very small amount (16.2%) of the variability in crude death rate. The model yields large p-values for all our variables, meaning that there is no statistical significance between our dependent variable and our indicators. This suggests that the indicators used may not be appropriate predictors for mortality trends, which are likely influenced by a more complex system of other factors such as conflict and/or disease outbreaks. This aligns with our exploratory data analysis, which showed high variability between crude death rate and internet access and no clear relationship. It is clear from our analysis that crude death rate is not easily predicted from socioeconomic factors alone.

Limitations:

Despite the statistical strength of our results, there are several methodological limitations which must be discussed. Our models assume linear relationships between indicators and dependent variables, which may not be the case. While we account for this by applying a log scale to our

GDP and health expenditure, the linear assumption may oversimplify the actual relationship between these features. Secondly, the adjusted r-squared and r-squared values are difficult to interpret without comparison to other models with different variables. Although we found a strong relationship between internet access and life expectancy, we may find better results when analyzing the relationship in conjunction with other indicators. Additionally, the low values found in our DPT immunization and crude death rate models indicate that other key variables may be missing. Lastly, we did not formally test for multicollinearity which could be influencing our results. There may be correlation between economic factors (GDP, Health Expenditure etc) which have not been accounted for in our study.

Future Works

Future studies can build upon our results in several ways. Incorporating additional variables related to healthcare, socioeconomics, geopolitics and technology can improve the accuracy and significance of our results. These new models could be tested, and their adjusted r-squared compared in order to see which combination of variables results in the highest score. Another option is a feature selection model, which after evaluating many factors and combinations could select the best model, and eliminate unnecessary variables. Formal tests for multicollinearity could also help in this regard, and further studies should examine the interaction between indicators. Secondly, further study should examine the strength of other non-linear models which may better capture the relationship between our chosen health metrics and internet access. More complex models may be needed to better explain the variance in our indicators. Lastly, future studies could look at time-series data to discover trends over time, and to account for any particularities from our chosen year (2015) which may be affecting our results.

References

[1] Anon. Metadata Glossary. Retrieved February 13, 2025 from https://databank.worldbank.org/metadataglossary/world-development-indicators/series/SP.DYN.CDRT.IN

[2] Byaro, Mwoya, et al. "Does internet use and adoption matter for better health outcomes in sub-Saharan African countries? new evidence from panel quantile regression." *Technological Forecasting and Social Change*, vol. 191, June 2023, p. 122445, DOI:https://doi.org/10.1016/j.techfore.2023.122445

[3] Gulshara Aimbetova, Gulzhan Dossybayeva, Roza Suleimenova, Aliya Turgambayeva, and Assiya Turgambayeva. 2022. Life expectancy as a quality of life indicator. *Iranian Journal of Public Health* (May 2022). DOI:http://dx.doi.org/10.18502/ijph.v51i5.9442

[4] Jessica Van Parys and Zach Brown. 2023. *Broadband internet access and health outcomes: Patient and provider responses in Medicare*. National Bureau of Economic Research Working Paper No. 31579 (August 2023). DOI:https://doi.org/10.3386/w31579

[5] Jiajie Yu and Shuang Meng. 2022. *Impacts of the internet on health inequality and healthcare access: A cross-country study*. *Frontiers in Public Health* 10 (June 2022). DOI:https://doi.org/10.3389/fpubh.2022.935608

[6] Kathryn E. Kerwin and Madison James. 2002. *The role of the internet in improving healthcare quality*. *Journal of Healthcare Management* 47, 4 (July 2002), 225–236. DOI:https://doi.org/10.1097/00115514-200207000-00005

[7] Moorhead, Jeffrey B., et al. "Internet access and cardiovascular death in the United States." *American Heart Journal Plus: Cardiology Research and Practice*, vol. 21, Sept. 2022, p.100200, DOI:https://doi.org/10.1016/j.ahjo.2022.100200

[8] Moussa, Ouattara Zieh, and Kenji Takeuchi. "Does searching online for vaccination information affect vaccination coverage? evidence from Sub-Saharan African countries." *Economics & Human Biology*, vol. 47, Dec. 2022, p. 101181, DOI:https://doi.org/10.1016/j.ehb.2022.101181

[9] The World Bank. 2024. *World Development Indicators*. Accessed February 2025. The World Bank Group. Available at: .https://datatopics.worldbank.org/world-development-indicators/

[10] United Nations, Department of Economic and Social Affairs, Population Division. 2015. *World Mortality 2015 Wall Chart*. United Nations.Available at: https://digitallibrary.un.org/record/831127/files/World_Mortality_2015_wall_chart_FINAL.pdf

[11] World Health Organization 2025 data.who.int, Diphtheria-tetanus-pertussis (DTP3) immunization coverage among 1 year olds (%) [Indicator]. https://data.who.int/indicators/i/48D7D19/F8E084C (Accessed on 29 January 2025)

AI Tool Declaration

AI was not used when completing this paper.