# Rethinking Health Care Systems: What are the Effects of Governmental/Social Health Insurance on Total Health Expenditure?

*Carla Garcia Medina*
*New York University*
*cgm396@nyu.edu*

*Mathilde Simoni*
*New York University Abu Dhabi*
*mps565@nyu.edu*

***Abstract*—National healthcare spendings vary a lot around the world. For instance, the US spends notably more than other similarly wealthy countries in Europe without better health outcomes. Studies have been conducted to explore the reasons for higher or lower costs for specific health services. However, little work has been done studying the relationship between healthcare expenses and healthcare systems. Different philosophies can be adopted: while some countries implement universal or governmental healthcare programs, others, such as the US, choose to mainly rely on private insurance. In this paper, we focus on the relationship between a country's type of healthcare system and its health expenses. The study was conducted for 42 countries using the Hadoop ecosystem (Pyspark and MapReduce). We found a negative correlation between the percentage of total population covered by governmental/social health insurance and the country's total health expenditure. These results were consistent over 5 years (2014 to 2018 included) and, with additional statistics, support that countries with governmental/social healthcare coverage spend less than countries with mostly private health programs. We hope that our results will help inform governmental organizations across the world when making decisions on how to restructure health care systems and reduce total health expenditure for both citizens and the government.**

***Keywords—Healthcare systems, Health expenditure, Data analysis, Correlation, Big data applications, Apache Hadoop, Apache Spark***

## I. INTRODUCTION

The rise of healthcare costs in recent years, and even before the COVID-19 pandemic, has highlighted the importance of ensuring the proper allocation of funds to healthcare resources. (WHO). While studies have been conducted to examine the effects of inner level aspects such as pricing, competition, and policy on total expenditure of a country, little work has been done on the effect of different health care insurance models in total health care expenditure.

According to the Columbia University Mailman School of Public Health, there exist four major types of health systems. The first one is known as the Beveridge Model, which is “both paid for and provided for by the government” and “free at point of service” (Columbia University Mailman School of Public Health). This system is adopted by the United Kingdom, New Zealand, Spain, and the U.S. Veterans Health Administration. The second major system is the Bismarck Model, also known as the social health insurance model in which healthcare is “paid for by non-profit insurance firms and provided by public and private actors” (Columbia University Mailman School of Public Health). This is the kind of model adopted in Germany, France, the Netherlands, Japan and Switzerland. Thirdly, there exists the National Health Insurance model that is in effect in countries like Canada, South Korea, as well as U.S. Medicare. This type of system is “paid for by government-run insurance programs and provided by public and private actors” (Columbia University Mailman School of Public Health). Finally, the Columbia University Mailman School of Public Health lists the Out-Of-Pocket model, known for having “health care paid for by consumers to public and private care providers” and “little to no insurance coverage.” This kind of model is in place in countries like Chad, India, and Rwanda.

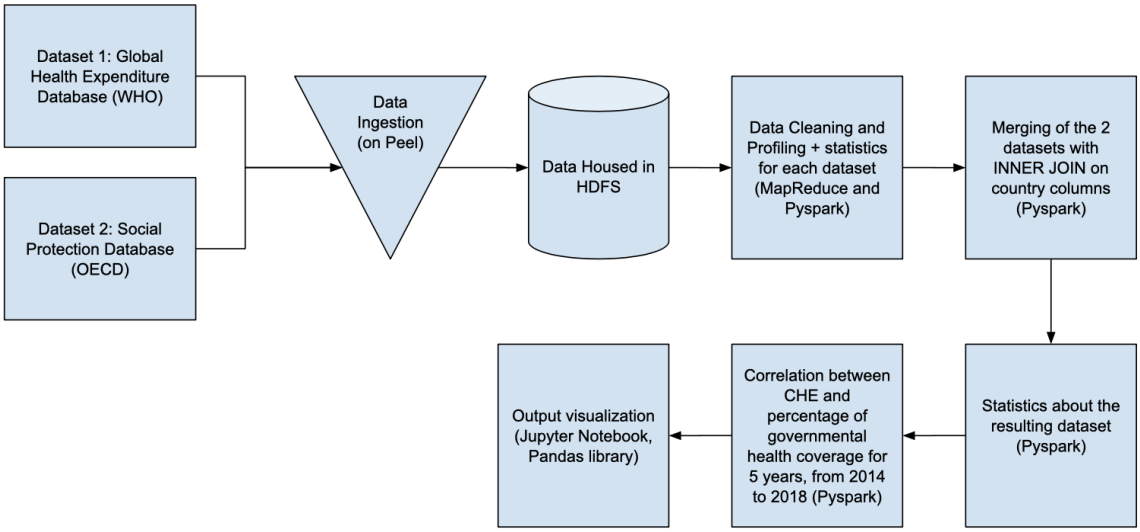


*Figure 1. Design Diagram*

This study aims to investigate the effects of the percentage of governmental/social health insurance in a country–defined by the OECD as the "share of population eligible for a defined set of health care goods and services under public programmes"–and its effect on total health expenditure. This study is hoped to benefit citizens and governmental organizations across the world, as the derived analytics could provide further insights into how to restructure health care systems to reduce total health expenditure. Both citizens and governmental organizations would reap the advantages of such modifications of health care systems as, by reducing total health care expenses, they would help reduce out-of-pocket expenses and/or contribute to the overall reduction of taxes.

In order to achieve this objective, the main goal is to calculate the correlation between the percentage of governmental/social health care of a country. Figure 1 shows how such a goal is to be achieved. As outlined in Figure 1, two datasets were employed in the study: one provided by the World Health Organization (WHO) containing information about the health expenditures in countries across the world, and the other provided by the Organisation for Economic Co-operation and Development (OECD) with data on the percentage of governmental/social insurance in all countries that belong to the organization. The two datasets were put on Peel and housed in HDFS. MapReduce and Pyspark were then used to do an initial cleaning and profiling of the individual data sources. Pyspark was then utilized to merge the two by performing an inner join based on the country column, as well as to further profile the merged dataset. The correlation between Current Health Expenditure and the Percentage of Governmental Health Coverage columns was then calculated for the years 2014 to 2018, inclusive. Finally, visualizations were output using the Pandas library in a Jupyter Notebook.

## II. MOTIVATION

The motivation behind this study lies in contradictory beliefs among citizens about the advantages of different types of healthcare systems. On the one hand, it can be claimed that if individuals have to pay for their own health care coverage, they would be more frugal regarding the health services they choose to ask for, avoiding unnecessary healthcare procedures and hence contributing to decrease the total national health expenditure. On the other hand, researchers Ryan Nunn, et al. in their article "A dozen facts about the economics of the US healthcare system" published in Brookings argue that the United States, known for not having a universal health care system, "spends more than other countries without obtaining better health outcomes." They add that this is the case for "surgical procedures, diagnostic tests, prescription drugs, and almost any other type of health-care service. " with "dramatically higher

health-care prices than other advanced countries." US healthcare expenses have also increased quickly, "rising from $2,900 per person in 1980 to $11,200 per person in 2018 (measured in 2018 dollars)—a 290 percent increase."

In addition, in their article for *The New York Times* Sarah Kliff and Josh Katzan explain how "insured patients [in the United States] are getting prices that are higher than they would if they pretended to have no coverage at all" because hospitals anticipate long negotiations with insurance companies. As a consequence, uninsured individuals appear to benefit from discounts while in reality, they are charged more than what an insurance would pay after negotiation. Thus, there seems to be a link between health care system types and total health expenditure.

As it has been and is still witnessed with the COVID-19 pandemic, decisions on healthcare and the population's general health is an important factor for a well-functioning society and economy in any country. In addition, Ryan Nunn, et al. highlight how excessively high expenses in healthcare can eventually necessitate "reduced spending on other important government functions like public safety, infrastructure, research and development, and education.. Therefore, it became necessary to study the differences in healthcare expenditure in distinct countries of the world, in particular its relationship with healthcare systems.

## III. RELATED WORK

No article was found that exactly studies the metric we measured but several studies compare health care expenditure of different countries and examine diverse causes for differences in expenses.

### *A.* ***"A dozen facts about the economy of the US healthcare system" by Ryan Nunn, et al.***

A notable observation in the analytic was that the US has the highest total health expenditure and the smallest percentage of population covered by governmental/social health insurance. Researchers Ryan Nunn, et al. confirm our observation and study diverse causes for that difference in health expenses. Some of them can be linked with the US healthcare system.

A first observation is that for privately insured patients, "health-care prices vary widely for the same service" and "some places […] have considerably higher health-care spending than others." However, for Medicare services, "prices are set administratively rather than through decentralized negotiations between payers and providers." This governmental program covers 34% of the population and leads to less variations in healthcare spending in the country. Hence, governmental coverage controls healthcare costs better and can help decrease the national health expenditure by not allowing physicians to choose prices based on negotiations with private insurance companies.

Moreover, Ryan Nunn, et al. mention that the high value for total healthcare expenditure in the US reflects rents ("payment to the health-care system beyond what is necessary for a normal rate of profit"). They are driven by the difficulty of patients paying to access quality health services and can be linked with the US healthcare system. Indeed, if the US had a more developed governmental/social health coverage, it would allow less out-of-pocket expenses, inducing less rents and a decrease in total healthcare expenditure.

The article also associates surprise billings with healthcare costs. Surprise billings are defined as "when insured patients find out […] that a provider […] was outside of their insurance network and is consequently much more expensive than they had anticipated." The writers mention that it "raises costs to consumers and allows providers to charge higher prices."

Finally, the researchers invoke other factors affecting the US total healthcare expenditure. They highlight that "administrative health-care costs are higher […] in the United States than in other countries." They also mention that "competition is unusually weak in the health-care system," which allows providers to raise the prices without losing patients. These factors must be included as possible causes for the high healthcare expenditure in the US and require further analysis.

### B. ***"Health care systems around the world" by Liji Thomas***

In this article, Dr. Liji Thomas compares different healthcare systems around the world and

supports our results that “Universal free healthcare is widely considered to be good for the country, health-wise as well as economically.”

In particular, she mentions that “the US healthcare system allows providers to inflate prices and expensive services, but poorly compensates essential services such as primary care and behavioral advice” supporting the analysis of Ryan Nunn, et al. in the first article.

When describing health systems in the UK and European nations, which are mostly based on universal/governmental free coverage paid by taxes except for special services handled by private insurances, she mentions that those “National health systems tend to control costs better” and “may increase the healthcare economy and efficiency.”

However, she also acknowledges some flaws in universal healthcare systems: “funding pressures are likely to go up as patients expect more advanced treatments as technology develops.” Governmental programs have strict policies for funding causing “the reluctance to recruit staff and to upgrade equipment.” On the contrary, “the US leads in medical innovation, boasting many of the world’s leading hospitals” and equipment can be funded by private actors more easily.

### *C.* ***“Projected costs of single-payer healthcare financing in the United States: A systematic review of economic analyses” by Cai et al.***

The authors performed this study due to growing interest in single-payer forms of healthcare (also known as the National Health Insurance Model) like *Medicare for All* in the United States. Although this type of program has gained public support, Cai et al observed that it fails to convince people that costs of health expenditure would be lowered, since people often associate it with unreasonable costs and regard it as an unsustainable system that forces governments to increase taxes.

The researchers compared cost analyses of 22 single-payer plans for the US, or individual states and all suggested long-term cost savings. The authors attribute the savings to lower drug costs and simplified billing.

Overall, the study achieved a “near consensus that single-payer plans would reduce expenditures while providing high-quality insurance to all US residents” but insist that that in order to “achieve net savings, single-payer plans [must] rely on simplified billing and negotiated drug price reductions, as well as global budgets to control spending growth over time” (Cai et al.). They conclude that “replacing private insurers with a public system is expected to achieve lower net health care costs.”

### *D.* ***“What is Universal Health Care?” by Mint Intuit***

This article provides a comprehensive description of Universal Health Care Systems, it’s subtypes (socialized medicine, single-payer system, and all payer systems) and outlines the advantages and disadvantages of universal healthcare systems in general, as well as of its specific subtypes.

In the article, it is stated that universal health care systems “lead to lower costs for both citizens and health care providers.” The authors explain that “under a universal health system, there is no competition between health insurance companies” and that “instead, the government regulates health care costs. This drives the cost of healthcare down substantially. Similarly, it reduces administrative costs for doctors and health care practitioners, as there is no need to deal with varying insurance companies.”

Further, the authors claim that “socialized medicine often is the least expensive, as the government has total and complete control over the cost of health.” While the authors do not list greater health expenditure in any of the disadvantages, they do mention that universal health care “may limit costly services that have low probability of success” and that universal health care “can take up an enormous portion of a government’s yearly budget.”

## IV. DESCRIPTION OF DATASETS

Two datasets were used in this study. The first one was obtained from the Global Health Expenditure Database, published by the WHO with link https://apps.who.int/nha/database/Select/Indicators/en. Its schema can be found at tinyurl.com/3dwb48wd. The dataset has a size of 16.6 MB (3649 rows and 2863 columns) and contains internationally comparable data on the yearly healthcare expenses of 192 countries from 2000 to 2018. The source contained data on different types of expenses for a variety of illnesses and specific healthcare domains.

| Field Name | Data type | Brief description |
|---|---|---|
| Country | String | 42 Countries (Europe, North and South America, as well as Australia, Japan, Israel) |
| Year | Integer | From 2010 to 2020 (2018 for some countries) |
| Health care coverage with **governmental/ social health insurance** | Float (% of total population) | Health care goods and services under public programs (share of population eligible to it) |
| Health care coverage with **private health insurance** (all types) | Float (% of total population) | |
| Health care coverage with *primary private health insurance* | Float (% of total population) | When there is no governmental/social coverage |
| Health care coverage with *duplicate private health insurance* | Float (% of total population) | Offers services already included under government health insurance but also offers access to different providers or levels of services |
| Health care coverage with *complementary private health insurance* | Float (% of total population) | Complements government/social insurance by covering all parts of the residual costs not otherwise reimbursed |
| Health care coverage with *supplementary private health insurance* | Float (% of total population) | Provides coverage for additional health services not covered by government/social insurance |

*Table 1. Dataset 2 Schema*

We mainly focused on analyzing the field Current Health Care Expenses per Capita in USD (CHE_PC_USD).

The second dataset was published by the Organisation for Economic Co-operation and Development (OECD) in the *OECD Health Statistics 2021*. The dataset has a size of 205KB (1611 rows and 11 columns) and the link to access it online is https://stats.oecd.org/Index.aspx?ThemeTreeId=9# (social protection tab). The dataset shows the percentage of total population covered by different types of health insurances in 42 countries, from 2010 to 2020 (or 2018 when later data was not available). A detailed schema of the health insurance 's types and a brief description for each column name and category can be found in table 1. In order to obtain information about the type of healthcare system in each country, we focused on the category *health care coverage with governmental/social health insurance.* A country with a high percentage has an important share of its population eligible to healthcare goods and services free under public programs, whereas a country with a low percentage mainly relies on private health insurances.

While the first dataset provides data for 190 countries, we restrained our analytic on the 42 countries for which health insurance coverage (type of healthcare system) data was available in the second dataset. We believe it did not significantly impact our study as it provided data for countries in various parts of the world including Europe, the Americas, Australia and Asia, as well as represented different types of economic systems.

## *V. ANALYTIC STAGES*

### *A.* ***Ingestion Process***

The two initial datasets *GHED_data.csv* and *dataset_initial.csv* were first downloaded from the

internet and added to the NYU Peel cluster using the scp command. Then, they were added into HDFS distributed file system to be able to perform initial cleaning and profiling.

### *B.* ***Dataset 1 Cleaning and Profiling Process***

An initial profiling and cleaning was done with two MapReduce jobs. As part of this initial profiling, a CountRecs job was first used to count the number of records in the original dataset. The number of rows output was 3648. The Clean job was then used to clean the data. All columns except for country and che_pc_usd were removed. Additionally, records with missing values in the che_pc_usd field were dropped. Only the values with information for the year 2018 were kept since this was the initial year of interest for the analytic. Running the CountRecs on the cleaned dataset (output in file with name part-r-00000), the number of rows output was 188. The new number of rows did not match the original one. The main reasons for this difference is that only rows with a year value of “2018” were selected since this was the initial year of interest for the analysis. Given that the data contained years from 2000 to 2018 inclusive, this resulted in a drastic reduction in the number of rows. Additionally, a few records had missing values for the che_pc_usd column. These records were removed since they were sparse and the countries in the dataset are extremely different to one another, such that imputing values could result in bias.

It was later decided to use PySpark to perform a deeper cleaning and profiling of dataset 1. (no use of the mapreduce job in the final product). The original dataset was used to calculate the covariance between Current Health Expenditure per Capita in US and subtypes of this expenditure including domestic general, domestic private, out-of-pocket, and primary health care. Descriptive statistics of these columns were obtained including the counts, means, standard deviations, minimum values, and maximum values. A list of the unique values for the countries column was output to see the individual countries present in the dataset. A list of the unique types of income groups present in the dataset was also shown. The covariance between each expense subtype and Current Health Expenditure was calculated.

The dataset was cleaned further with Pyspark such that the column names were renamed to improve readability and usability of the dataset (e.g. "income group (2018)" was renamed to "income_group"). The data types of the six main health expenditure columns were casted to integer to be able to perform mathematical operations on these. Only the "country", "income_group", "year", "che_pc_usd" columns were selected for the cleaned version of dataset1 since these would be the columns of interest for merging and the analytic. The data frame was filtered to only keep data for the last five years present in the dataset (2018 to 2014, inclusive). All remaining records with NULL values were shown and since these formed a very small minority of the total records in the dataset, they were dropped. The name of the “country” column was changed to “countries” so that it would not coincide with the name of the "countries" column in dataset2. The names of some countries were renamed to match the names of countries in dataset2 so that merging the two datasets would be done correctly. The final dataset was then output to obtain a general idea of what it looked like.

### *C.* ***Dataset 2 Cleaning and Profiling Process***

MapReduce was used for an initial cleaning of dataset 2. The file was first filtered to keep data for the years 2014 to 2018. Then, it was reorganized to contain 8 columns: "country", "year", "Total health care", "Total private", "Primary private", "Duplicate private", "Complementary private" and "Supplementary private". The last 6 columns collect percentages of total population for different types of healthcare coverage in 42 countries between 2014 and 2018. Finally, missing values were replaced with the string "NONE".

It was later decided to use PySpark to perform profiling and a deeper cleaning of dataset 2. Descriptive statistics for 2018 were obtained including the counts, means, standard deviations, minimum values, and maximum values for each type of health insurance. Countries with the minimum and maximum values were also displayed. Finally, column names were renamed to improve readability and usability of the dataset (e.g. *"_c3" was renamed to "total_private_coverage")* and values were casted to float to be able to perform mathematical operations on these.

When profiling the initial dataset, we realized that there were too many missing values for all health

coverage columns except for governmental_coverage. Initially, the method called LOCF (Last Observation Carried Forward) was examined. It consists of replacing any missing value with the last value observed in the previous years. However, this strongly assumes that the value of the outcome remains unchanged over time and the manipulation was impossible for many inputs as the data was mostly missing for all the years before the one studied. Hence, only the "country", "year", "governmental_coverage" columns were selected for the cleaned version of dataset2. Governmenttal_coverage was the only column that would give consistent values and was of interest for merging and the analytic. The few remaining missing values for this field were dropped as they formed a very small minority of the total records in the dataset (leading to 39 countries for the merging of 2018).

### *D.* ***Merged Dataset Cleaning and Profiling Process***

The correlation between the che_pc_ud column in Dataset 1 and the governmental_coverage column in Dataset 2 was calculated for each year between 2014 and 2018. First, the two dataframes were filtered to only keep data for the year being analyzed. Then, they were merged in a new dataframe using the pyspark inner join function based on the country name. Finally, the correlation was calculated using the pyspark corr function. Additional statistics on the distribution of the two columns, including the count, mean, standard deviation, minimum and maximum values, were also calculated for every year. Tables containing the che_pc_usd for the countries with minimum and maximum governmental coverages were also output.

## VI. GRAPHS - VISUAL REPRESENTATIONS OF THE ANALYTICS

### *A.* ***Dataset 1***

Figure 2 shows the descriptive statistics about the main types of expenses present in dataset 1. Figure 3 includes box plots to visualize these distributions. The abbreviation PC stands for Per Capita and USD refers to the fact that the values are representative of the cost in US dollar

| summary | che_pc_usd | gghed_pc_usd | pvtd_pc_usd | oop_pc_usd | phc_usd_pc |
|---|---|---|---|---|---|
| count | 3579 | 3001 | 3389 | 3531 | 200 |
| mean | 868.0569991617771 | 174.3615461512829 | 178.35379167896136 | 160.5202492211838 | 214.925 |
| stddev | 1519.5288265681004 | 225.45497608349336 | 232.82943856553896 | 214.49178969751554 | 233.62719814603744 |
| min | 4 | 0 | 0 | 0 | 10 |
| max | 10624 | 995 | 992 | 999 | 992 |

*Figure 2. Descriptive Statistics Table of Main Types of Expenses in Dataset 1*

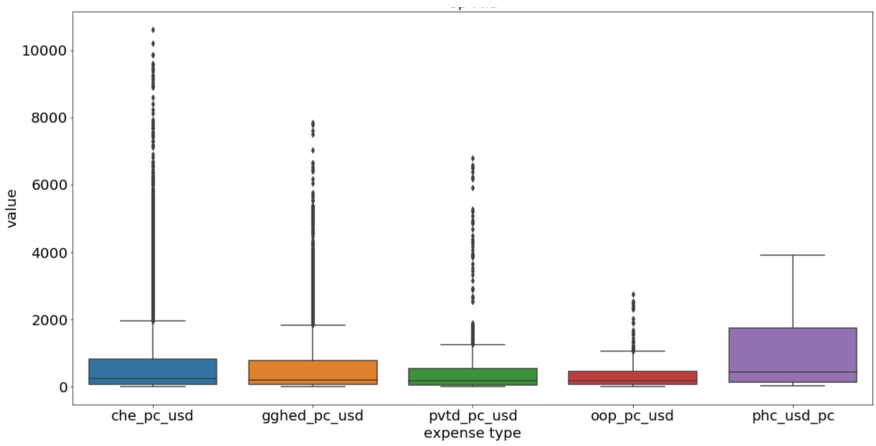


*Figure 3. Box Plots of Distributions of Main Types of Expenses*

The types of expenses are abbreviated as follows:

- CHE: Current Health Expenditure
- GGHED: Domestic General Government Health Expenditure
- PVTD: Domestic Private Health Expenditure
- OOPS: Out-of-Pocket Expenditure (OOPS)
- PHC: Primary Health Care Expenditure

In Figures 2 and 3, it can be observed that the distributions are positively skewed, that there are many outliers, and similar spreads for all expenses except for phc_usd_pc, which has a significantly greater standard deviation. The positively skewed distribution suggests that most countries tend to have lower health care related costs, while the large number of outliers, particularly for the che_pc_usd column might reveal that there are a few countries whose health care costs are much greater than those of the rest of the world. Finally, noting that the similar spread and lower variance of expenditure types is suggestive that most countries have similar health expenditures that are not too different from the mean. Nevertheless, the greater variation in phc_usd_pc could be indicative of a greater variety of opinion regarding the importance of allocating funds to primary health care.

Figure 4 contains descriptive statistics about the distribution of the ratios of the aforementioned expenditure types with che_pc_usd to analyze what types of expenditure contribute the most to total expenditure. Figure 5 consists of box plots that visually represent these statistics. These two figures illustrate how there is a normal distribution, similar spreads for all expenses, and the GGHED:CHE ratio is the closest to 1, suggesting a stronger relationship, followed by PHC, then PVTD, and finally, OOPS. This would imply that the greatest contributor to the total expenses, in general, is the Domestic General Government Health Expenditure .

| summary | gghed_over_che | pvtd_over_che | oop_over_che | phc_over_che |
|---|---|---|---|---|
| count | 3001 | 3389 | 3531 | 200 |
| mean | 0.46209984481045296 | 0.4132064403078131 | 0.33989509105208215 | 0.5396500481146899 |
| stddev | 0.20973454124551666 | 0.1991811973944592 | 0.19822505831125212 | 0.1326003320778862 |
| min | 0.0 | 0.0 | 0.0 | 0.3147502903600465 |
| max | 0.9944289693593314 | 1.0 | 1.0 | 0.9259259259259259 |

*Figure 4. Descriptive Statistics Table of Ratios of Expenditure Types with che_pc_usd in Dataset 1*

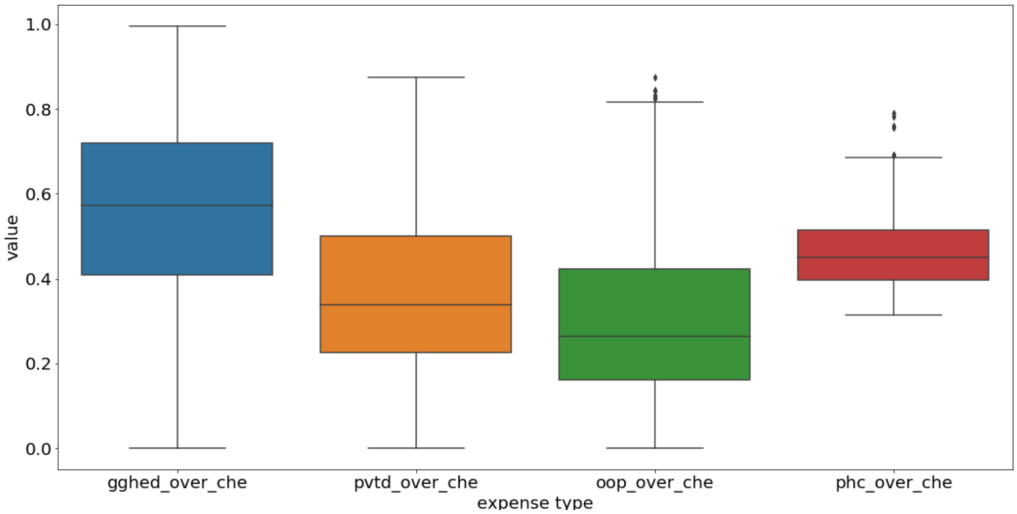


*Figure 5. Descriptive Ratios of Expenditure Types with che_pc_usd in Dataset 1*

Figure 6 shows the covariance between the other expenses and CHE for countries classified as low income. This income group was the one with the least amount of data points and, therefore, we do not see results for the oop expenditure. However, it is possible to appreciate that that GGHED has the greatest covariance with expenditure, revealing that in low income countries GGHED is the variable that is most strongly related to CHE, PVTD has a negative regression, indicating a negative relationship with PVTD. In other words, Domestic Private Health Expenditure, would appear to reduce the total health expenditure in low income countries.

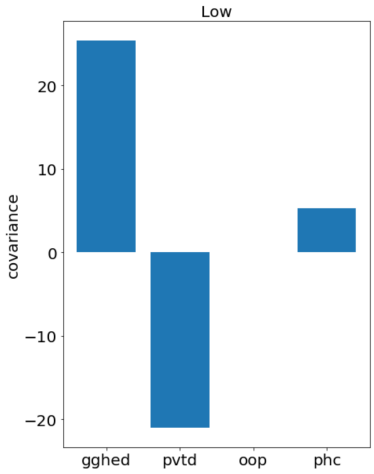


*Figure 6. Covariance between Other Types of Expenses and CHE in Low income countries*

Figures 7, 8, and 9 also show that GGHED is also the type of expenditure with the strongest covariance, or relationship, to the CHE in low-middle income countries, high-middle income countries and high income countries. Notwithstanding, while the covariance is smallest for PVTD in Figures lower middle and higher middle income countries, this is not the case in high income countries, where it is out-of-pocket expenditure that appears to contribute the least to the total health expenditure of a country.

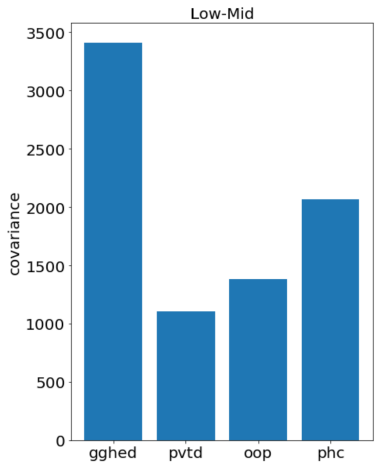


*Figure 7. Covariance between Other Types of Expenses and CHE in Lower Middle Income Countries*

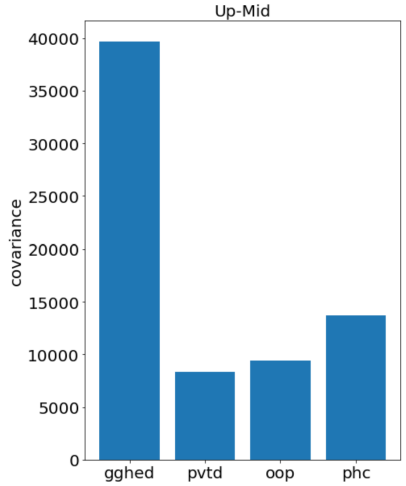


*Figure 8. Covariance between Other Types of Expenses and CHE in Upper Middle Income Countries*

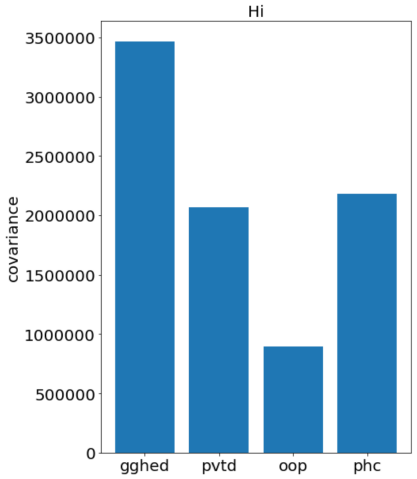


*Figure 9. Covariance between Other Types of Expenses and CHE in High Income Countries*

Figure 10 illustrates the distribution of CHE across income categories for the year 2018. These results were extremely similar to those of all years from 2014 to 2018 inclusive, suggesting a negligible variation in CHE in the last five years present in the dataset. It can be appreciated that total health care expenses are substantially greater for high income countries. Another aspect worth noting is that the CHE increases as the income category becomes higher.

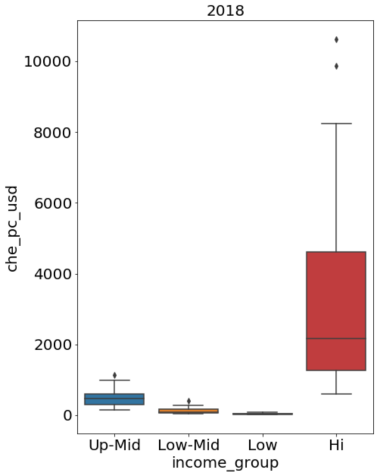


*Figure 10. Distribution of che_pc_usd across income categories for the year 2018*

*B.* ***Dataset 2***

Figure 11 shows the descriptive statistics about the main types of healthcare coverage in dataset 2. Figure 12 includes box plots to visualize these distributions. It can be observed that the mean for the percentage of total population covered by governmental/social insurance is very high (95.98%). However, the mean for the 5 other types of private health insurance is below 33%. This suggests that the majority of countries in dataset 2 have a healthcare system with a strong dependence on governmental/social insurances.

```
---------
VARIABLE: governmental_coverage
---------
- mean value: 95.98205116467598

- minimum value: 34.0
countries with the minimun value:
+-------------+
|      country|
+-------------+
|United States|
+-------------+

- maximum value: 100.0
countries with the maximum value:
+--------------+
|       country|
+--------------+
|     Australia|
|        Canada|
|Czech Republic|
|       Denmark|
|       Finland|
|        Greece|
|       Ireland|
|        Israel|
|         Italy|
|         Japan|
|         Korea|
|        Latvia|
|    Luxembourg|
|   New Zealand|
|        Norway|
|      Portugal|
|      Slovenia|
|         Spain|
|        Sweden|
|   Switzerland|
|United Kingdom|
+--------------+
```

*Figure 13. Countries with governmental/social coverage minimum and maximum values*

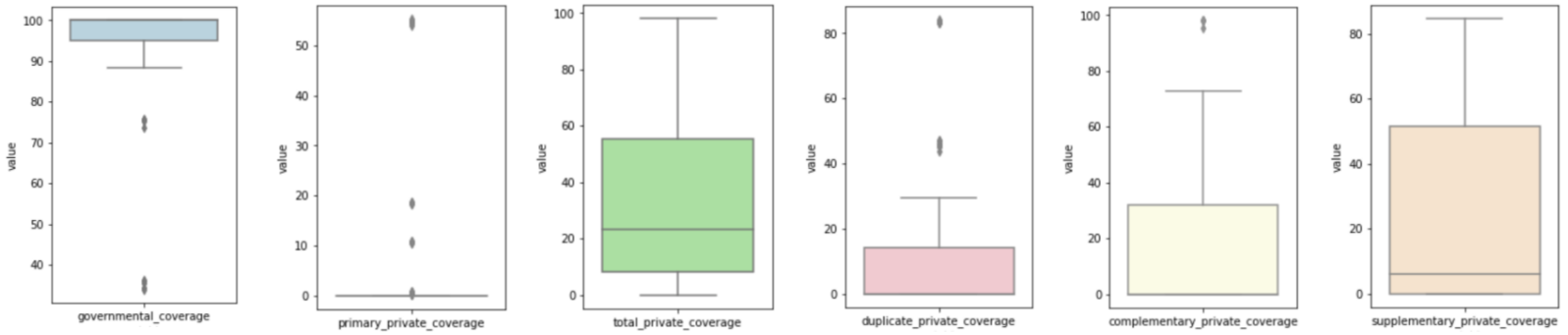


*Figure 11. Box Plots of distributions of Healthcare Coverage Types in Dataset 2*

| summary | governmental_coverage | total_private_coverage | primary_private_coverage | duplicate_private_coverage | complementary_private_coverage | supplementary_private_coverage |
|---|---|---|---|---|---|---|
| count | 39 | 29 | 24 | 21 | 23 | 20 |
| mean | 95.98205116467598 | 32.72758638961562 | 3.48333336537083 | 13.023809387570335 | 20.108696009801783 | 25.840000057220458 |
| stddev | 11.290584687547835 | 30.73377508910274 | 11.648275400904762 | 21.83311453908841 | 30.545300874688994 | 31.64354745500076 |
| min | 34.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| max | 100.0 | 98.0 | 54.5 | 84.1 | 98.0 | 84.1 |

*Figure 12. Descriptive Statistics Table of Healthcare Coverage Types in Dataset 2*

Figure 13 shows the countries with maximum and minimum percentage for governmental/social healthcare coverage. We can observe that the range of values is large. The United States has the minimum percentage (34%) of its population covered by a governmental/social program, while 21 countries, mostly in Europe as well as Australia and New Zealand, Canada, Israel and north Asia have 100% of their population covered by governmental/social healthcare coverage.

### *C. **Merged Dataset***

Figure 14 shows descriptive statistics about the main columns of interest in the merged dataset. Regarding the che_pc_column, the mean total expenditure is about 3700. This value could be of interest to individuals who purchase private insurance and can compare their current expenditure to the mean of the countries in the resulting dataset. The standard deviation is relatively large at almost the same order of magnitude as the mean. Finally, the range just like for governmental coverage is extremely large, ranging from 390 to 10624.

| summary | governmental_coverage | che_pc_usd |
|---|---|---|
| count | 39 | 39 |
| mean | 95.98205116467598 | 3677.3076923076924 |
| stddev | 11.290584687547835 | 2612.91639999452 |
| min | 34.0 | 390 |
| max | 100.0 | 10624 |

*Figure 14. Descriptive Statistics of govenmental_coverage and che_pc_usd Columns in the Merged Dataset.*

The correlation between governmental_coverage and che_pc_usd was calculated for every year from 2015 to 2018. The results are shown in Figure 15. The correlation ranges from -0.16 to -0.26. The fact that the correlations are similar and negative indicate that increasing the percentage of governmental coverage could help reduce total health expenditure of a country.

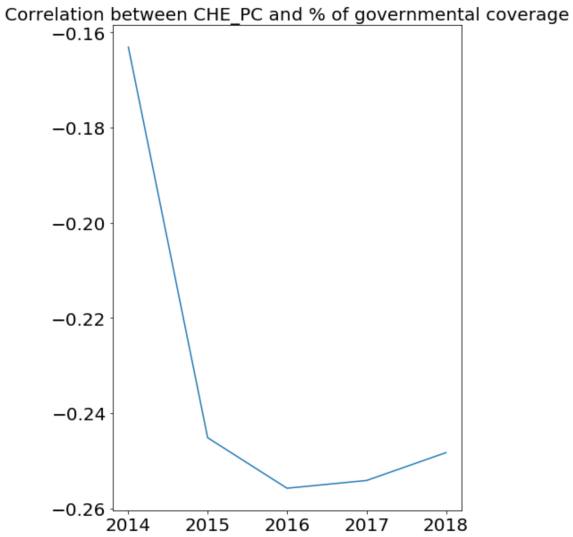


*Figure 15. Correlations between govenmental_coverage and che_pc_usd Columns from 2014 to 2018.*

In order to investigate explanations for these correlations, we looked at the che_pc_usd values for countries with the lowest and the highest percentage of governmental_coverage. Figure illustrates the results for the minimum percentage of governmental coverage for the year 2018. In this case, the country with the least percentage of governmental coverage is the United States at 34%. It's expenditure is 10624. Because these two values represent the minimum governmental_coverage value and maximum che_pc_usd as previously seen in Figure 14, there is reason to believe that the United States may be having a large influence in the final correlations obtained. Similar results were observed for all years between 2014 and 2018.

```
Minimum value for government insurance coverage: 34.0
Countries with the minimum value for governmental
insurance coverage and their total health expenditure:
+-------------+------------+----------+
|      country|income_group|che_pc_usd|
+-------------+------------+----------+
|United States|          Hi|     10624|
+-------------+------------+----------+
```

*Figure 16. Che_pc_usd Values for countries with Lowest Percentage of Governmental Insurance*

Figure 17 shows the countries with the highest percentage of governmental coverage. In this case, 100%. A large number of higher income countries can be seen, mostly from Europe and Asia.

These have varied yet more moderate, che_pc_usd values. The results were relatively similar for years from 2014 to 2018. It is possible that these findings suggest that having universal health insurance might lead to more moderate results. However, it would be worth further investigating if the US has such large che_pc_usd values because of confounding variables, not present in the dataset that are unrelated to governmental coverage.

```
Maximum value for government insurance coverage: 100.0
Countries with the maximum value for governmental
insurance coverage and their total health expenditure:
+--------------+------------+----------+
|       country|income_group|che_pc_usd|
+--------------+------------+----------+
|        Canada|          Hi|      4995|
|Czech Republic|          Hi|      1766|
|       Denmark|          Hi|      6217|
|       Finland|          Hi|      4516|
|        Greece|          Hi|      1567|
|       Ireland|          Hi|      5489|
|        Israel|          Hi|      3324|
|         Italy|          Hi|      2989|
|        Latvia|          Hi|      1101|
|    Luxembourg|          Hi|      6227|
|        Norway|          Hi|      8239|
|      Portugal|          Hi|      2215|
|      Slovenia|          Hi|      2170|
|         Spain|          Hi|      2736|
|        Sweden|          Hi|      5982|
|   Switzerland|          Hi|      9871|
|United Kingdom|          Hi|      4315|
|     Australia|          Hi|      5425|
|         Japan|          Hi|      4267|
|   New Zealand|          Hi|      4037|
|         Korea|          Hi|      2543|
+--------------+------------+----------+
```

*Figure 17. Che_pc_usd Values for countries with Highest Percentage of Governmental Insurance*

## VII. CONCLUSION

In this paper, we studied the relationship between a country's type of healthcare system and its total health expenditure. Using two datasets provided by the World Health Organization (WHO) and the Organization for Economic Co-operation and Development (OECD), we found a negative correlation, consistent over 5 years, between the percentage of population covered by governmental/social health insurance and the country's total health expenditure. These results indicate that countries with a well-developed governmental/social healthcare system spend less than countries mainly relying on private health insurances.

As discussed in this paper, some reasons for this outcome may be that government-controlled healthcare systems rely on simplified billing, help reduce out-of-pocket expenses and rents, and properly allocate funds to healthcare resources, leading to a decrease of the national health expenditure. Additionally, they create a more equal system by regulating drug and health service prices, which allows less variations for health expenses across the country.

The global rise of healthcare costs, the impact of the COVID-19 pandemic on healthcare systems' efficiency and the debates about the structure of the US healthcare system in recent years show that new reforms are needed for more efficient and less wasteful healthcare systems. Our work shows that restructuring healthcare systems with governmental/social programs could help decrease national health expenditure. We hope it will inspire and help governmental organizations when making these decisions.

However, correlation doesn't imply causality. Other factors such as innovation and less competition have been identified as possible causes for high healthcare costs in the US and need to be deeper analyzed. Furthermore, we observed that most of the countries in the datasets we used have a healthcare system with a strong dependence on governmental/social insurances, except a few, such as the US. Thus, our analytic could be biased. Future works need to study lower income countries and a larger number of data points.

## VIII. ACKNOWLEDGEMENTS

We would like to thank NYU High Performance Computing (HPC) for giving us access to NYU Peel cluster, PySpark and Hadoop tutorials as well as for their support. The assistance provided by Professor Ann Malavet by answering our questions and giving valuable advice was also greatly appreciated. Finally, we must thank the World Health Organization (WHO) as well as the OECD (Organization for Economic Co-operation and Development) for providing us access to the 2 datasets we used in our analytic.

## IX. REFERENCES

Cai, Christopher et al. “Projected costs of single-payer healthcare financing in the United States: A systematic review of economic analyses.”

*PLoS medicine* vol. 17,1 e1003013. 15 Jan. 2020, doi:10.1371/journal.pmed.1003013

Chernew, Michael E., et al. "Reducing Health Care Spending: What Tools Can States Leverage?" *Commonwealth Fund*, https://www.commonwealthfund.org/publications/fund-reports/2021/aug/reducing-health-care-spending-what-tools-can-states-leverage.

Columbia Mailman School of Public Health. "Types of Health Systems." Search the Website, https://www.publichealth.columbia.edu/research/comparative-health-policy-library/types-health-systems-0.

Global Health Expenditure Database, World Health Organization, https://apps.who.int/nha/database/Select/Indicators/en.

Nunn, R., Parsons, J., & Shambaugh, J. (2020, April 6). A dozen facts about the economics of the US health-care system. Brookings. Retrieved October 13, 2021, from https://www.brookings.edu/research/a-dozen-facts-about-the-economics-of-the-u-s-health-care-system/.

OECD. OECD Statistics, https://stats.oecd.org/Index.aspx?ThemeTreeId=9#.

Princeton Public Health Review. "Health Care Reform: Learning from Other Major Health Care Systems." *Princeton University*, The Trustees of Princeton University, https://pphr.princeton.edu/2017/12/02/unhealthy-health-care-a-cursory-overview-of-major-health-care-systems/.

Roy, Baijayanta. "All about Missing Data Handling." *Medium*, Towards Data Science, 27 June 2021, https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184.

Thomas, Dr. Liji. "Healthcare Systems around the World." *News*, 6 Apr. 2021, https://www.news-medical.net/health/Healthcare-Systems-Around-the-World.aspx.