Final Report Research Paper

a. Introduction (Topic Background, Research Questions, and Impact):

i. Background

In the United States, health expenditures have grown dramatically over the past several decades, outpacing general economic growth. National health spending is driven by multiple factors including technological advances, rising costs of treatments, an aging population, and changes in healthcare financing. At the same time, public health outcomes—measured by metrics such as disease-specific mortality—are influenced by a wide range of factors such as medical innovation, public health interventions, and social determinants of health.

ii. Research Questions:

Our research aims to explore the relationship between these rising health expenditures and the burden of disease, as captured by the IHME Global Burden of Disease (GBD) dataset. Specifically, we aim to answer the following questions:

1: How do overall national health expenditures (both in total and per capita) relate to trends in mortality, as measured by IHME data?

2: How does the composition of health expenditures—broken down by service types (e.g., hospital care, physician services) and financing sources (e.g., government, private sponsors)—impact mortality from specific diseases?

iii. Impact

- **Policy Implications:** Both questions have strong implications for healthcare policy. If we find that increased spending is associated with improved mortality outcomes, or that certain types of expenditures are particularly effective, this information can guide future budget allocations and policy reforms.
- **Efficiency and Equity:** Investigating the composition of spending can reveal whether resources are being used efficiently and equitably. For example, if public spending in preventive care correlates strongly with lower mortality in chronic diseases, it might suggest that expanding such programs could lead to better outcomes.
- **Long-Term Trends:** With healthcare costs rising steadily over the past decades, understanding the relationship between spending and health outcomes is critical. These research questions help to contextualize historical trends and to predict future challenges, especially as the population ages and new medical technologies emerge.

Overall, our research questions address both the macroeconomic perspective—how total spending relates to mortality—and the more granular perspective—how the allocation of spending across different services influences specific health outcomes. This dual approach provides a comprehensive understanding of the healthcare spending landscape and its impact on public health.

b. Related work/ work that has been done in this area by others

- **David M. Cutler – "Is Technological Change in Medicine Worth It?"**
    - Argues that technological advances are a primary driver of rising healthcare costs.
    - Finds that these innovations often lead to substantial improvements in health outcomes (e.g., better survival from cardiovascular events and neonatal care).
    - Relevant to our project because it raises the question: is increased spending translating into better outcomes across all diseases, or only some?
- **GBD 2019 – "Global burden of 369 diseases and injuries in 204 countries and territories…"**
    - Provides comprehensive data on mortality and morbidity trends from 1990–2019.
    - Offers the disease-specific death rates we use in our analysis.
    - Validates the importance of looking at which diseases have improved (or worsened) despite spending increases.
- **Gerard F. Anderson et al. – "It's the Prices, Stupid: Why the United States Is So Different from Other Countries"**
    - Shows that the U.S. spends more on health care not because of overuse, but due to higher prices for services, drugs, and administrative costs.
    - Supports the motivation for asking: Is the U.S. getting what it pays for in terms of mortality reduction?
    - Emphasizes that high spending doesn't necessarily equate to better outcomes.
    - References (ACM Format)

[1] David M. Cutler. 2004. Is Technological Change in Medicine Worth It? Health Affairs 23, 2 (2004), 11–16. DOI: https://doi.org/10.1377/hlthaff.23.2.11

[2] GBD 2019 Diseases and Injuries Collaborators. 2020. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. The Lancet 396, 10258 (2020), 1204–1222. DOI: https://doi.org/10.1016/S0140-6736(20)30925-930925-9)

[3] Gerard F. Anderson, Uwe E. Reinhardt, Peter S. Hussey, and Varduhi Petrosyan. 2003. It’s the Prices, Stupid: Why the United States Is So Different from Other Countries. Health Affairs 22, 3 (2003), 89–105. DOI: https://doi.org/10.1377/hlthaff.22.3.89

c. Data

i. Datasets

We are using the following datasets, converted from csv into SQL:

- IHME GBD 2021 dataset (Global Burden of Disease Study), which contains yearly estimates of disease burden and mortality for specific causes in the United States.
- NHE (National Health Expenditure) Tables from the Centers for Medicare & Medicaid Services (CMS), including:
- Table 1: Total Expenditures
- Table 2: Expenditures by Type
- Table 3: Source of Funds
- Table 5: Type of Sponsors
- Table 23: Real Dollar Adjustments (inflation-adjusted expenditures)

ii. Why these particular datasets?

We selected these datasets because they are:

- Authoritative and comprehensive, coming from trusted sources (IHME and CMS).
- Time-aligned, providing historical annual data from 1980 to 2023 (with some exceptions).
- Complementary, allowing us to link mortality data from IHME with financial expenditure patterns from NHE to investigate trends and possible relationships.
- These datasets together provide a rich foundation for analyzing how healthcare spending relates to public health outcomes over time.

iii. Data cleaning steps:

We performed the following cleaning steps:

1. **Dropped irrelevant or redundant columns** in the IHME dataset (e.g., location_id, age_id, sex_name) that contained only one value or were not meaningful for U.S.-level aggregate analysis.

2. **Converted year columns to rows** (unpivoted) in NHE tables using pandas melt() to standardize formats across all datasets.
3. **Merged inflation data** from Table 23 to transform nominal expenditures to real 2017-dollar values.
4. **Dropped tables that lacked necessary temporal range or were duplicative**, such as Table 4 and Table 19.
5. **Standardized** column names for merging consistency.
6. **Handled non-numeric symbols** like ' - ' by replacing them with NaN and converting relevant columns to numeric types.
7. **Verified absence of missing values** in critical columns after cleaning.

These steps ensured that all datasets could be reliably joined on the year field and used for consistent analysis across all research questions.

iv. Some cautions about the data

- The IHME dataset contains mortality estimates (val, upper, lower) expressed as counts, not rates. While useful, they do not adjust for population growth.
- Some NHE tables begin in 1987, meaning that not all tables cover the full 1980–2023 range. This affects which years can be used in merged analyses.
- Missing or placeholder values like ' - ' were present in some tables and needed to be replaced with NaN.
- National Health Expenditure totals in the real-dollar tables are aggregates of other rows and should be excluded from comparative category analyses.
- A few disease-specific trends from IHME show unexpected spikes or drops, which might be due to changes in measurement or classification. These should be interpreted cautiously.

Overall, the datasets were of high quality, and after cleaning, we were confident in their integrity and usability for the purposes of this research.

d. Methodology:

i. Methodology

To answer our two research questions, we followed this structured approach:

- **Data Integration and SQL-Based Querying**
    - All cleaned and transformed datasets were loaded into a relational SQL database.
    - We designed a schema that connects all tables using year as the primary linking key.
    - SQL was used to:
        - Perform joins across datasets.
        - Select relevant disease-specific rows.
        - Aggregate expenditure totals by year and subcategory.
        - Filter and reshape data for further analysis in Python.
        - Python-Based Visualization and Analysis

- **Python (with pandas, matplotlib) was used to:**
    - Calculate descriptive statistics.
    - Identify skewed distributions and trends.
    - Plot time-series trends for diseases and expenditures.
    - Compare nominal vs. real dollar expenditures.
    - Visualize top-5 expenditure categories and most prevalent diseases.
    - Explore potential relationships between expenditures and death counts.

- **We also used Python for data wrangling tasks that would be cumbersome in SQL, such as:**
    - Converting strings to numeric values.
    - Plotting disease-specific death trends year-over-year.

- **Correlation Analysis**
    - To explore whether higher spending correlates with reduced mortality (first research question), we merged real-dollar health expenditure data with total death counts from IHME.

        - We examined scatter plots and computed Pearson correlation values for an initial assessment.

ii. Why this methodology?

- **We chose SQL for the majority of the work because:**
    - The datasets were structured and relational in nature.
    - SQL excels at filtering, joining, grouping, and aggregating tabular data efficiently.
    - It's a core requirement for this project and enforces discipline in data design and normalization.
- **We used Python for the remaining because:**
    - It's more expressive and flexible for data cleaning, numerical conversions, and plotting.
    - Python's libraries like pandas and matplotlib make time-series visualizations and ad-hoc analyses much easier.
    - Tasks like detecting and replacing placeholder characters (e.g., ' - ') or drawing grouped plots would be tedious in SQL.

This SQL-first and Python-assisted approach allowed us to maintain data integrity, leverage the strengths of both tools - statistically summary, aggregations along with complex analysis capability, and ensure low margin for error when evaluating trends and relationships.

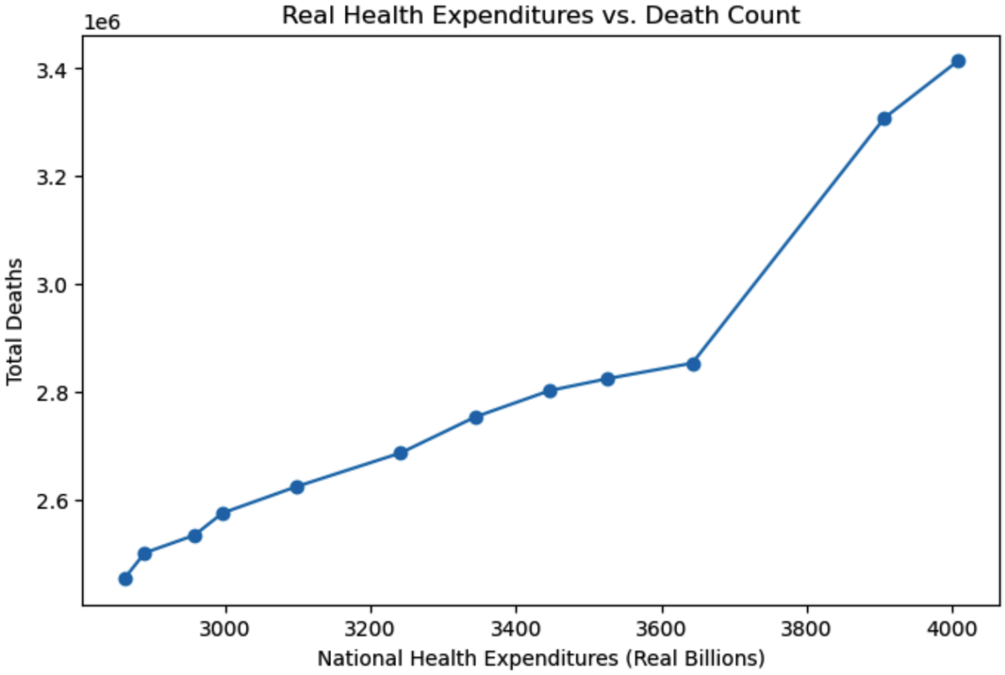
e. Results and Discussion

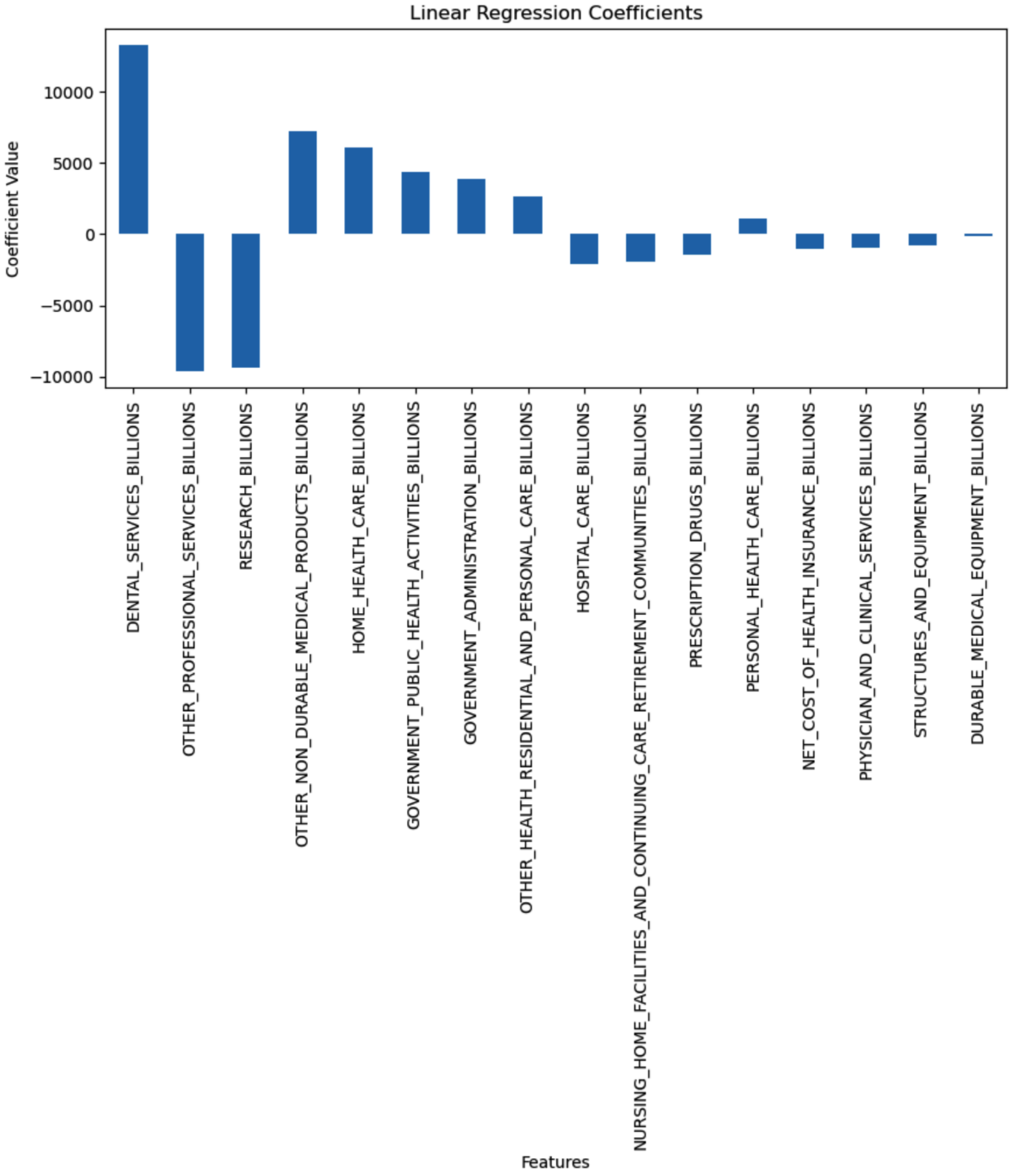
i. Results

- **Research Question 1:** How do overall national health expenditures (both in total and per capita) relate to trends in mortality, as measured by IHME data?
    - There’s a clear positive linear relationship between mortality rate and national health expenditures. Diving deeper into the specifics, although there’s fluctuations in the mortality rates between the disease, the type of disease doesn’t correlate to expenditures, just the mortality rate does.

- **Research Question 2:** How does the composition of health expenditures—broken down by service types (e.g., hospital care, physician services) and financing sources (e.g., government, private sponsors)—impact mortality from specific diseases?
    - Tight interweaving of rising disease burden and growing healthcare expenditures. However, not all spending categories are equally impactful on outcomes like death rates.

ii. Visualizations:





iii. Validity

1. Internal validity:

- Strengths
  - The data used comes from authoritative, structured datasets such as NHE and IHME → consistent tracking over time.

- Comprehensive analysis incorporating: Correlation analysis, linear regression for feature importance, and trends in spending and funding sources/sponsors
- Clear patterns emerged (e.g., strong positive correlation between deaths and medical service/product spending), aligning with the hypothesis that healthcare expenditure dynamics can explain some mortality trends.

- Weaknesses:
  - **Correlation is not causation:** the fact that spending and deaths rise together doesn't mean one causes the other.
  - Factors such as lifestyle, social determinants of health, environmental influences, and medical innovations were not included in this analysis.

2. External validity:

- Strengths
  - The data spans over 40 years → long-term trends
  - Data is from national-level sources → enhanced generalizability within the U.S.
  - Visualizations and regression models support the general perception that rising costs are borne increasingly by institutions, and that healthcare is growing more complex and expensive.
- Weaknesses
  - The findings may not generalize outside the U.S., where healthcare systems differ drastically in structure.
  - Some data (e.g., deaths) may be influenced by external shocks (e.g., pandemics) or reporting inconsistencies, which aren't explicitly modeled.
  - The effect of policy, efficiency, and preventive measures is not isolated, so the insights may not fully reflect real-world impact evaluations.

f. Discussion of results:

- **Cases where you do not have meaningful results**
  - Our analysis was limited to national-level data, without stratifying by age, gender, region, or policy periods, which could provide more nuanced insights.
- **Impact on the community**
  - Policy makers and healthcare planners can interpret this as an indication that increasing mortality rates lead to systemic cost increases, potentially motivating preventive care investment.
  - Researchers might use this finding to argue for more disease-specific funding analyses or to promote investments that reduce mortality burden through upstream interventions.
- **Limitations that are not discussed in validity:**
  - The temporal relationship between health spending and mortality may involve time lags (e.g., spending now may reduce future mortality), which wasn't captured in a linear model.
  - All analysis was at the national level, which hides inequalities across states, ethnic groups, or age categories.

G. Future work:

- **How can people use these results?**

    - In addition to what is discussed in the impact on the community part above, Economists and data scientists studying healthcare systems can use this as a baseline correlation model for resource allocation or efficiency analyses.

- **If more time were given:**
    - Move beyond correlation to causal inference by applying techniques like time-lagged regression, Granger causality tests, or structural equation modeling to better understand if and how spending influences mortality (or vice versa).