A Deep Dive Into Factors Impacting Healthcare Affordability

Vivek Kumar Bhagat

High drug prices in the U.S. strain many Americans, making healthcare unaffordable for low- and middle-income families. In my project about the question, "How Does the Pharmaceutical and Insurance Industry's Self-serving Goals Impact Healthcare Affordability In The United States Of America?" I intend to analyze the key drivers behind high drug prices and their impact on healthcare costs.
Using publicly available datasets on drug prices, healthcare spending, and income levels, I intend to explore patterns in pricing strategies, regional cost differences, and the effect of transparency laws on prices. I'll also analyze insurance roles, and pharmaceutical monopolies to uncover where price escalations happen and how they affect access to necessary medications.

1. *QUESTION*

How Does the Pharmaceutical and Insurance Industry's Self-serving Goals Impact Healthcare Affordability In The United States Of America?

2. *DATA SOURCES*

The analysis of this project relies on two comprehensive datasets that provide complementary perspectives on healthcare affordability in the United States. The **CMS Prescription Drug Utilization and Spending Data [1]** offers detailed insights into prescription drug spending, claims, unit costs, and utilization patterns within Medicare and Medicaid programs, shedding light on the economic burden of pharmaceutical pricing and its impact on patient access. Meanwhile, the **National Health Expenditures (NHE) Tables [2]** present a broad view of U.S. healthcare spending trends, segmented by service type, funding sources, and historical changes. Together, these datasets enable a nuanced exploration of how drug pricing policies and broader healthcare spending patterns intersect, offering a deeper understanding of affordability challenges in the healthcare system.

2.1. Data Structure

The CMS Prescription Drug Utilization and Spending Data is organized into temporal, categorical, and numerical variables. Temporal data, such as the year, allows for an extended analysis of drug spending patterns. Categorical variables include drug names, and prescriber identifiers, among others, providing dimensions for comparative analysis. Numerical variables capture total and average spending amounts, total claims, and unit costs, offering a detailed financial and usage perspective. Although the dataset is robust, certain fields have missing entries, which were addressed during the data preprocessing stage to enhance usability.

The National Health Expenditures (NHE) Tables dataset features a combination of temporal, categorical, and quantitative data elements. Temporal variables, such as years, track healthcare spending over time, highlighting trends and shifts in expenditures. Categorical variables classify the data by healthcare service types (e.g., hospital care, prescription drugs) and funding sources (e.g., government programs, private insurance). Quantitative variables include expenditure totals and percentage changes, offering insight into spending patterns across different sectors. The dataset, while extensive, requires cleaning and consolidation due to its multi-sheet Excel format, ensuring it can be seamlessly integrated into the analysis pipeline.

2.2. Data Quality

The **CMS Prescription Drug Utilization and Spending Data** demonstrates strong data quality overall but has a few areas for improvement. In terms of **accuracy**, the dataset reflects real-world Medicare and Medicaid spending trends, with drug claims and spending figures aligning with published CMS reports. However, its **completeness** is slightly compromised by missing values in certain cells which required imputation during preprocessing. The data shows good **consistency**, as fields such as spending amounts and claims are uniformly formatted. Regarding **timeliness**, the dataset is updated annually, making it suitable for examining historical trends and current patterns, though it may lag slightly behind rapidly evolving policy changes. The most recent data is from 2022, which was used for this project. Finally, its **relevancy** is high, as the data directly supports the project's objective of analyzing drug pricing and access within federal insurance programs. Descriptive statistics revealed expected distributions in spending values, and outliers were addressed by removing unrealistic entries, such as weighted averages for certain drugs.

The **National Health Expenditures (NHE) Tables** dataset also maintains high data quality, particularly in its **accuracy**, with expenditure values verified against official government reports. The dataset is **complete** for its intended use, providing a comprehensive breakdown of healthcare spending by service type and funding source. **Consistency** across files was addressed during preprocessing, as minor formatting inconsistencies were identified and corrected, such as differing column headers across Excel sheets. **Timeliness** is appropriate, with data spanning multiple decades, making it ideal for analyzing long-term trends, though it may not fully capture the most recent changes in spending patterns. The dataset is highly **relevant** for evaluating broader healthcare expenditure dynamics. Automated validation checks **confirmed the absence of invalid or duplicate values**, and descriptive statistics provided insights into spending trends and variability, ensuring the data's readiness for integration and analysis.

2.3. F.A.I.R. Guidelines

CMS Prescription Drug Utilization and Spending Data

The dataset is publicly provided under U.S. government guidelines, adhering to the **Open Government Data Act (2018) [3]** and public domain principles. It allows unrestricted use, modification, and redistribution as long as no false claims of endorsement by CMS are made. No explicit attribution is legally required by CMS but is considered a best practice.

FAIR Guidelines:

Findability: The dataset is hosted on the CMS data portal with a well-documented API and metadata for easy discovery.

Accessibility: Delivered in a user-friendly CSV format, it is accessible to anyone with internet access and basic data processing tools.

Interoperability: The standard CSV format ensures high compatibility with data analysis platforms, and the use of universally understood schema names further supports its usability.

Reusability: The data's public domain nature and detailed documentation enhance its reusability, provided users include proper attribution when applicable.

National Health Expenditures (NHE) Tables

This dataset, like the CMS data, is also under the U.S. government guidelines, adhering to the **Open Government Data Act (2018) [3]** as it is a product of a U.S. government agency. There are no restrictions on its use, reuse, or redistribution, and it can be freely incorporated into academic and research projects.

FAIR Guidelines:

Findability: The dataset is well-archived and indexed on the CMS website with direct download links for ease of access.

Accessibility: Available in a compressed ZIP file containing Excel spreadsheets, it is accessible via common tools but requires initial unpacking for analysis.

Interoperability: The data is stored in Excel files, which are widely supported, but the lack of uniform formatting across sheets requires preprocessing to ensure smooth interoperability.

Reusability: The public domain status ensures that the dataset can be reused across projects with no legal or licensing restrictions, provided its structure and content are properly understood.

3. DATA PIPELINE

The data pipeline was implemented using Python, leveraging libraries such as pandas, requests, sqlite3, and zipfile for data extraction, transformation, and loading. It was designed to automate the process of acquiring, cleaning, and preparing the two datasets for analysis, ensuring efficiency and reproducibility.

The pipeline begins by downloading the datasets directly from their respective sources. For the **CMS Prescription Drug Utilization and Spending Data**, the file is retrieved via a static API request and loaded into a pandas DataFrame. For the **National Health Expenditures (NHE) Tables**, the ZIP file containing multiple Excel sheets is downloaded and extracted, data is then cleaned and transformed to ensure consistency and usability.

Transformations and Cleaning Steps:

<u>CMS Prescription Drug Utilization and Spending Data</u>

1. Dropped unnecessary columns to focus on key metrics such as drug spending, claims, and unit costs, reducing noise in the data.

2. Standardized column names for better consistency and ease of analysis.

3. Converted elements of column *Brand Name* into proper format for further analysis.

<u>National Health Expenditures (NHE) Tables</u>

1. Extracted and concatenated relevant Excel sheets into a unified pandas DataFrame.

2. Ensured consistent column naming and removed redundant headers caused by multi-sheet formats.

3. Picked required cells while dropping the remainder to maintain a consistent format for easier manipulation.

4. RESULTS AND LIMITATIONS

The data pipeline successfully processed and consolidated the CMS Prescription Drug Utilization and Spending Data and the National Health Expenditures (NHE) Tables into clean, structured formats ready for analysis. For the CMS dataset, unnecessary columns were removed, missing geographic data were imputed, and numeric fields were standardized, resulting in a concise dataset focused on drug spending trends, claims, and unit costs. The NHE dataset was cleaned and unified from multiple Excel sheets, resolving header inconsistencies and consolidating data on healthcare spending by funding source. Together, these outputs provide a robust foundation for exploring the relationships between drug utilization, healthcare expenditures, and policy impacts.

Additionally, the datasets reflect only annual updates and do not capture real-time changes, reducing their suitability for analyzing recent policy shifts, thus the scope of the pipeline focuses on descriptive data preparation and does not yet include advanced integration or modeling steps, which limits the depth of insights it can provide. These constraints highlight areas for improvement in the pipeline's flexibility and scalability for future analysis.

5. REFERENCES

1. Centers for Medicare & Medicaid Services. (n.d.). CMS Prescription Drug Utilization and Spending Data. Retrieved from https://data.cms.gov/data-api/v1/dataset/87604795-a3e2-4190-9b3a-e39142221fcd/data.csv

2. Centers for Medicare & Medicaid Services. (n.d.). National Health Expenditures (NHE) Tables. Retrieved from https://www.cms.gov/files/zip/nhe-tables.zip

3. U.S. Chief Information Officers Council. (n.d.). Open, Public, Electronic, and Necessary (OPEN) Government Data Act (OGDA). Retrieved from https://www.cio.gov/handbook/it-laws/ogda/