Medicare Provider Performance, Patient Demographics, and Service Impact

G. Brint Ryan College of Business, University of North Texas

DSCI5260: Business Process Analytics

Group - 9

Ajay Sai Balaji Sunkari

Bysani Alekya

Mounika Gavvala

Ravi Chandrika Yarramreddy

Sarvamangala Sahithi Pulugurta

Acknowledgements

We deeply appreciate the G. Brint Ryan College of Business at the University of North Texas for establishing the academic base and resources needed to complete this project.

Our research received essential support from Professor, Dr. Sameh Shamroukh, who provided guidance and feedback and encouragement throughout the research process. Your business analytics and healthcare data expertise guided us through every research phase while providing essential direction and clarity.

The Centers for Medicare & Medicaid Services (CMS) receives our appreciation for their public release of Medicare data which allowed us to study real-world healthcare issues effectively.

We thank our teammates Ajay Sai Balaji Sunkari, Bysani Alekya, Mounika Gavvala, Ravi Chandrika Yarramreddy, and Sarvamangala Sahithi Pulugurta for their dedication, collaboration, and commitment. The project success depended on each member's distinctive work in data analysis and modeling as well as literature review and documentation.

We express our gratitude to our families and peers who provided continuous support during the entire duration of this project.

Abstract

The research provides an extensive evaluation of Medicare provider performance together with patient demographics and chronic disease prevalence and an assessment of how drug-based services compare to direct medical services for patient outcomes. The research uses CMS Open Data portal analytics to address healthcare inefficiencies and demographic disparities and rising costs of chronic disease management.

The project uses Business Intelligence tools together with statistical testing and machine learning models such as Random Forest Regression and K-Means Clustering to analyze more than 1.2 million rows of provider-level data. The analysis methods allow researchers to detect top-performing providers while grouping patients by demographics and assessing treatment success across various healthcare delivery methods.

The analysis shows that higher Medicare payments and increased service delivery typically lead to better patient results yet these connections remain inconsistent which indicates payment system inefficiencies and potential misalignments. The research shows major differences between geographic areas and population groups because older patients and those from underserved communities receive inferior medical results.

The research demonstrates fundamental distinctions between drug-based and medical services regarding their cost-effectiveness which requires more sophisticated resource allocation methods. The research develops a data-driven framework through data preprocessing and feature engineering and regression modeling and clustering to support healthcare policy development and promote equity and operational decision-making within the Medicare ecosystem. The research adds value to academic literature about healthcare analytics while providing practical guidance to policymakers and providers who want to improve Medicare services through effectiveness and equity and sustainability.

Executive Summary

This research presents an in-depth evaluation of Medicare provider performance, with a focus on the influence of patient demographics, chronic conditions, and the impact of drug-based versus direct medical services on healthcare outcomes and cost efficiency. The study utilizes the Centres for Medicare & Medicaid Services (CMS) Open Data Portal, analyzing over 1.2 million rows of Medicare Part B data to uncover key insights aimed at informing policy and improving healthcare strategies.

Using advanced machine learning techniques, statistical tests, and data visualization tools, the study examines critical patterns in provider performance, demographic disparities, and treatment outcomes. The results reveal that while higher Medicare reimbursements and service volumes typically correlate with improved patient outcomes, these relationships are not universally strong. Inefficiencies in the current payment models were identified, with certain high-cost providers showing limited improvements in patient outcomes.

Significant regional and demographic disparities were also uncovered, particularly concerning aging and minority populations, who often face limited access to quality care. The findings highlight the need for policy reforms that focus on addressing these disparities and improving efficiency in the Medicare system.

Key research findings include:

- High-performing Providers: Identified through clustering and regression analysis, these providers tend to deliver better outcomes for high-risk patients but show inefficiencies in cost allocation.
- Disparities in Access: Aging and minority groups experience worse access to care and treatment outcomes, necessitating targeted policy interventions.

- Cost and Outcome Trade-offs: A nuanced understanding of how different treatment models impact both clinical and fiscal results, with a focus on improving cost-efficiency.

Recommendations for Policy and Strategy include:

1. Incentivizing High-Performing Providers: Encourage best practices through value-based payment models and performance incentives.
2. Targeted Interventions for Demographically Disadvantaged Groups: Develop policies to improve care access for underserved communities.
3. Optimizing Resource Allocation: Reevaluate reimbursement models to better align cost with patient outcomes, ensuring efficient care delivery.

This study provides actionable insights aimed at enhancing the effectiveness, equity, and sustainability of Medicare services, with implications for future healthcare policy and strategic planning.

2.Introduction

Millions of Americans, especially the elderly and those with chronic conditions, receive care from the U.S. healthcare system, which is a significant role in supporting Medicare-supported services. However, there are still many challenges in ensuring the effectiveness of the provider's work, the equity of patient care, and the optimal use of resources.

2.1 Current healthcare operations often struggle with:

- Variations in provider performance: Healthcare providers' efficiency and patient satisfaction are high while others have operational bottlenecks and low performance scores.
- Demographic disparities in care delivery: Patient outcomes and service utilization are further influenced by socioeconomic factors, regional healthcare access, and demographic influences.
- High chronic disease prevalence among Medicare beneficiaries: Healthcare costs are higher due to the existence of chronic diseases including diabetes, heart disease and COPD and the diseases are also costly because they are usually long-term conditions that require management.
- Cost differences between drug-based and medical services: Medical services that receive Medicare patients are either pharmaceutical based treatment or direct medical intervention but the effect of these on patient outcomes and resource efficiency is not well understood.
- This research tries to uncover patterns in provider performance, patient demographics, chronic disease prevalence and treatment outcomes by leveraging Business Intelligence (BI) and Data Analytics.

2.2 Project Goals and Tools to be Used

In this project, data analytics and Business Intelligence (BI) tools will be used to analyze Medicare provider performance, demographic disparities, chronic disease patterns and the impact of different treatment approaches.

2.2.1 Primary Objectives

- Provider Performance Analysis

  Discover high performing Medicare providers by payment allowances and reimbursement amounts. Figure out the correlation between the provider's performance and patient satisfaction scores. Reveal whether financial incentives affect provider efficiency and service quality.

- Patient Demographics and Disparities

  A study of patient demographic distributions by age, sex, location, and socioeconomic status across different geographic regions is required. The existence of potential healthcare disparities is identified by demographic variables. Explore if certain patient groups get varying levels of care or whether there are any delays in treatment.

- Chronic Condition Prevalence

  Analysis of chronic diseases among Medicare beneficiaries. Explore if the prevalence of chronic diseases differs by provider or geographic region. Review of how different providers tackle chronic diseases and reveal the best approaches.

- Impact of Drug and Medical Services

  Study the patient's outcomes based on whether they get drug-based treatments or direct medical services. Examine the cost differences between pharmaceutical management and the actual medical care. Find out the effectiveness of the resources utilization in

patient's health outcomes during the management process under the different treatment models.

2.2.2 Proposed Tools and Technologies:

To achieve these objectives, this project will utilize the following tools: Python (Pandas, Seaborn, Scikit-learn, Matplotlib): Statistical analysis, trend forecasting, and visualization. Machine Learning Algorithms: Predictive modeling for provider performance and treatment effectiveness. Tableau: Interactive dashboards visualization.

2.3 Research Questions

2.3.1 Provider Performance Analysis:

Who are the high performing providers dependent on Medicare payments and the amounts allowed? What is their correlation to patient satisfaction?

2.3.2 Patient Demographics and Disparities:

What is the variation in patient demographics (age, sex, etc.,) spread across various geographies and is there any disparity identified because of the demographics across various geographies?

2.3.3. Chronic Condition Prevalence:

Taken medicare beneficiaries, identify the distribution of chronic diseases. Does it vary in way with the provider in place or by geographic location?

2.3.4. Impact of Drug and Medical Services:

What is the anticipated patient outcome based on the service (drug or medical) service offered by the provider? Is there any differences between these two services in costs or resource allocation?

3.Literature Review

The present-day health care system is filled with inefficiencies that start from improperly allocated resources, extremely limited access to care and unanticipated dissatisfaction among the patients and their families. These bottle necks are very evident in the Medicare chain where the services are always in demand, which is spurred by the age group of beneficiaries who have specific chronic conditions, which keep growing with every passing day. To resolve these persisting issues and pain points, data driven approach is the best solution that not just analyses performance of various providers in the system but also proposes optimised approaches that reshape the impact of healthcare services based on various patient demographics.

In this literature review, we aim to synthesize recent contributions that are appreciated and implemented. As a part of this review, we identify gaps and limitations, throw a light on the key contributions, methodologies applied and outline how this body can pave the way for future investigations. Taking these contributions as foundation blocks, this research aims at contributing to an equitable and efficient health care system.

3.1 Care and patient satisfaction

3.1.1 Overview: In their work, Perals, Rambaud, and García (2024) have focused on the relationship between the quality of care offered to patients and their satisfaction which throws light on the economic effects in the healthcare system. This study brings out the need to have a patient centric care that does not only improve satisfaction but also uncovers the influence of demographic factors and provider performance.

3.1.2 Methodology: The authors used a mixed-methods approach of convergent parallel design, which integrated quantitative analysis of patient satisfaction surveys with qualitative interviews of healthcare providers. Thematic analysis was used to interpret

provider perspectives, and regression models were used to identify predictors of patient satisfaction.

3.1.3 Gaps & Limitations: The study has some limitations; the study is based on aggregated data, and this may hide disparities at the individual or provider level, and the study does not fully explain the impact of chronic conditions or geographic variations on patient satisfaction.

3.1.4 Propositions & implications for this research: Expanding on the work of Perals et al., this research will use granular, provider-level data to analyze the relationship between provider performance and patient satisfaction in conjunction with demographic and geographic factors.

3.2 Payment approaches and quality of chronic care

3.2.1 Overview: To disentangle the impact of alternative payment models (APMs) and their associated service delivery models on the quality of chronic care, Simmons et al. (2024) conducted a scoping review. They found that APMs can improve care quality but also exacerbate disparities in access and outcomes for vulnerable populations.

3.2.2 Methodology: A systematic scoping review was employed in the study, analyzing 45 studies published between 2010 and 2023. Thematic analysis was used to synthesize the data, and the findings were categorized by payment model type, patient population and outcomes.

3.2.3 Gaps and Limitations: The review found that there is limited research examining the interaction between APMs and patient demographics and provider performance to affect the management of chronic diseases. Moreover, many studies included in the review are based on self-reported data, which may be biased.

3.2.4 Propositions & implications for this research: Utilising this approach, the current research will fill these gaps by exploring how Medicare payment systems affect

provider performance and patient outcomes for particular emphasis on the chronically ill population.

3.3 Provider Performance & Roles

3.3.1 Overview: In their study, Patel et al. (2023) examined the use of evaluation and management (E/M) visits by nurse practitioners (NPs) and physician assistants (PAs) in the USA from 2013-2019. They determined that NPs and PAs are gradually assuming the role of delivering E/M services, especially in underserved regions.

3.3.2 Methodology: A cross-sectional time series design was used in the study analyzing Medicare claims data to monitor fluctuations in E/M visits by NPs and PAs. Multivariate regression models were employed to determine the impact of provider type on service utilization and costs.

3.3.3 Gaps and Limitations: These findings raise concerns about how the transition to NP and PA E/M service delivery models affects patient care and healthcare expenditures. Moreover, the study does not fully address how patient characteristics influence these outcomes.

3.3.4 Propositions & implications for this research: Building on the work of Patel et al., this research will extend their analysis to investigate how provider performance and patient satisfaction relate to resource allocation, with emphasis on demographic factors.

3.4 Medicare and Financial Protections

3.4.1 Overview: In their paper, Chernew and Masi (2024) explained the challenges of balancing access to care and financial protections within Medicare as the program ages. Future reforms must address disparities in access and outcomes while also maintaining fiscal sustainability, they suggest.

3.4.2 Methodology: The study used a policy analysis framework to review existing literature and Medicare policy documents to identify key challenges and make recommendations.

3.4.3 Gaps and Limitations: A gap in the knowledge on how Medicare payment structures affect provider performance and patient outcomes of the beneficiaries with chronic conditions was also established.

3.4.4 Propositions & implications for this research: This gap will be filled by this research which will analyze the effect of Medicare payment structures on provider performance and patient outcomes using detailed provider level data.

3.5 Outcomes of Value-Based Purchasing for Hospital

3.5.1 Overview: In their studies, Banerjee et al. (2019) and Lee et al. (2019) investigated the effect of Medicare's Hospital Value-Based Purchasing (HVBP) program on hospitals' performance and patient outcomes. Banerjee et al. (2019) argued that there was a positive relationship between the level of exposure to HVBP and the reduction in 30-day mortality rates; Lee et al. (2019) identified enhanced hospital operational results.

3.5.2 Methodology: Banerjee et al. (2019) compared hospitals with different levels of exposure to HVBP using a difference-in-differences approach, while Lee et al. (2019) analyzed operational outcomes using econometric models.

3.5.3 Gaps and Limitations: Both studies have limited ability to account for patient demographics and geographic disparities; thus, more nuanced analyses are needed to understand how value-based programs affect different populations.

3.5.4 Propositions & implications for this research: These studies will be expanded upon by the research proposed here, which will explore how provider performance and patient outcomes are affected by value-based programs, with attention to demographic and geographic factors.

3.6 Inequities in care – An overview

3.6.1 Overview: McMurtry et al. (2019) and Timbie et al. (2020) labored to define the disparities of the healthcare system particularly in availability and provision of services

to specific racial and ethnic groups. McMurtry et al. (2019) reported that Asian Americans are discriminated within healthcare systems and Timbie et al. (2020) concluded that differences are present in care provided between health system-affiliated and non-affiliated physician institutions.

3.6.2 Methodology: McMurtry et al. (2019) analyzed experience frequencies of discrimination based on a self-report approach whereas Timbie et al. (2020) compared the quality of care by physician organization types based on regression models.

3.6.3 Gaps and Limitations: These studies nonetheless lack the evaluation of the provider's performance or payment models in analyzing the influence of demographic factors and systemic biases on access to care and patient outcomes.

3.6.4 Propositions & implications for this research: The gaps will be filled by this study, which will explore the way provider performance and demographics interact to establish disparities in care.

Briefly, this research explores gaps found in literature by attempting:

It will give a comprehensive evaluation of patient satisfaction and provider performance. It will also compare the chronic disease burden by region and among providers. It assesses the impact of specific services on patient outcomes and costs. It will yield actionable results to eliminate the gap, enhance utilization of resources and make healthcare services more effective in the Medicare system.

This study will also educate a more nuanced understanding of the determinants of inefficiencies and variation in healthcare delivery by using a rich dataset and sophisticated analytical methods and thereby inform policy and practice translating into improved patient outcomes.

4. Research Method

Medicare Physician & Other Practitioners by Provider - CMS Open Data

4.1 Data source:

The CMS (Centers for Medicare & Medicaid Services) Open Data Portal offers detailed, standardized and accurate datasets on Medicare providers, services and payments. The dataset chosen for this study is appropriate for our purpose because it contains essential components that are necessary to assess provider performance, patient demographics, chronic disease incidence and service effects.

4.2 Key Reasons for Choosing This Dataset:

4.2.1. Provider Performance Analysis

- Includes Medicare payment details (allowed amounts, submitted charges and reimbursements) which can be used to measure provider efficiency and financial reliance on Medicare.
- Providers provide level data that can be used to identify high performing providers by reimbursement trends and correlations with patient outcomes.
- Enables evaluation of regional differences in provider performance by location and specialty.

4.2.2. Patient Demographics and Disparities

- It gives information on the demographics of the providers based on their locations and the kind of services they provide.
- It also helps in the analysis of the availability of health care by relationship between distance and population distribution in combination with demographic patterns.

4.2.3. Chronic Condition Prevalence

- Enables comparison of the services delivered for conditions, and monitoring the patterns of chronic disease management
- Assists in determining if the prevalence of chronic diseases differs by provider or area and provides information on how healthcare resources are distributed.

4.2.4. Impact of Drug and Medical Services

- Includes information on the kinds of services and treatments (CPT codes) which can be used to compare pharmaceutical therapy with medical management.
- Enables cost-benefit analysis by determining the reimbursement rates, patient costs, and the effectiveness of the services.
- Help in identifying the drug versus medical service utilization patterns and thus in helping in informing policy decisions.

Advantages of Using CMS Data

- Official and Reliable: CMS is the main source of Medicare data, and therefore accurate and reliable.
- Very detailed Provider Level Data: The dataset is able to provide information on a individual provider level which enables comparison of the performance.
- Comprehensive Coverage: All Medicare enrolled physicians and practitioners in the U.S. are included, which provides a wide yet structured scope.
- Available to the Public: There are no legal or administrative restrictions to the data, which is useful for research.

4.3 Dependent Variable (Y Target Variable):

`Bene_Avg_Risk_Scre` (Average Patient Risk Score)

This variable is an aggregate measure of the overall health risk of Medicare beneficiaries, based on the co-occurrence of chronic conditions and their impacts on providers. It is thus an ideal metric for assessing patient outcomes as a function of provider type, service utilization and healthcare costs.

Correlates with all research questions:

Provider Performance: Increased payment or service frequency could be related to improved patient outcomes.

Patient Demographics & Disparities: The risk score is different for various demographics. Chronic Condition Prevalence: More chronic diseases increase the risk score.

Impact of Drug & Medical Services: Various treatments affect the mean risk score.

4.4 Independent Variables (X Predictors)

The following independent variables hold a strong impact on `Bene_Avg_Risk_Scre`:

4.4.1 `Rndrng_Prvdr_Type` (Provider Type MD, NP, PA) :

Variety of providers (Medical Doctor (MD), Nurse Practitioner (NP), Physician Assistant (PA)) use varying treatment strategies. May manage more chronic or complicated illnesses, whereas NPs or PAs may concentrate on disease prevention and management. It can be used to classify healthcare disparities by provider type.

4.4.2 `Tot_Srvcs` (Total Services Provided)

A direct effect on patient results: A sign of better patient care: More services offered by a provider means:

More patients → More experience.

More number of treatments → Better healthcare.

Providers who provide more services may have better patient risk management and therefore better results.

4.4.3 `Tot_Mdcr_Pymt_Amt` (Total Medicare Payment Amount)

A key financial indicator of not only the optimal utilization of resources but also the efficiency and effectiveness of providers. It is possible that higher Medicare payments are linked with better quality of care and more sophisticated services. It can be used to help analyze cost differences among providers and regions.

4.4.4 `Bene_CC_PH_Hypertension_V2_Pct` (Hypertension Prevalence in Beneficiaries)

Hypertension is a leading predictor of many chronic diseases, affecting patient health scores.

The higher prevalence means that more attention should be paid to this population, which will lead to higher service utilization and costs. A strong predictor of long-term patient outcomes.

4.4.5 `Bene_Avg_Age` (Average Age of Beneficiaries)

- Age is one of the most important predictors of health risks.
- Older patients have more chronic diseases and thus need more health care.
- It can be used to help identify demographic disparities in risk scores.

4.5 Technical Analysis Methods

Different models were selected to ensure that the model chosen is accurate, easily interpretable and not costly to compute.

- Random Forest Regressor (Stable, NonLinear, Missing Values Accepted)
- K-Means Clustering

4.6 Assumption Testing:

The reliability of any statistical modeling or inference should be checked by testing key assumptions related to normality and linear relationships between variables. Two tools were used:

- Correlation Matrix
- Shapiro-Wilk Normality Test

4.6.1 Correlation Matrix (Linearity Assumption)

The correlation matrix checks the strength of linear relationships between numerical variables.

The correlation between tot_srvcs and tot_benes amounts to 0.34 which indicates a moderate positive relationship. The relationship between total services and total beneficiaries shows a tendency to increase together.

The variables bene_avg_age and bene_avg_risk_scre demonstrate weak or negligible correlations with all other variables since their values approach 0. The data shows no strong linear relationship between these variables.

Interpretation:

The analysis confirms that multicollinearity does not present a significant issue in this particular dataset. The weak correlations indicate that linear models have limitations in relationship capture so additional complex models should be considered.

Pair	Correlation Coefficient	Interpretation
tot_srvcs & tot_benes	0.34	Moderate positive linear relationship. As the number of services increases, the number of beneficiaries increases.
tot_srvcs & bene_avg_age	0.14	Very weak correlation. Little to no linear association.
tot_srvcs & bene_avg_risk_scre	-0.10	Very weak negative correlation. Practically negligible.
Others	~0.0 to 0.2	All weak or no significant linear relationship

Table 1: Correlation Matrix Details

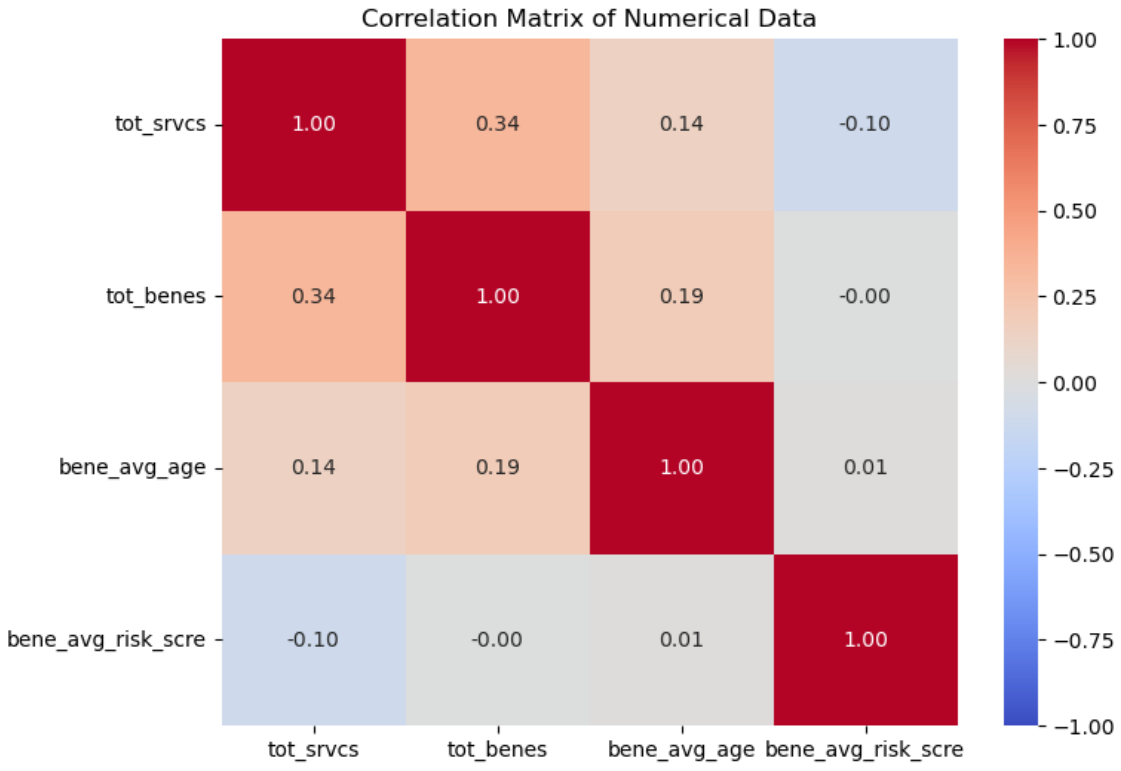

Fig 1: Correlation Matrix

4.6.2 Shapiro-Wilk Normality Test (Normality Assumption)

Purpose: The test determines whether a variable follows a normal (bell-shaped) distribution which many statistical tests and machine learning models require as an essential assumption.

Used before parametric tests (like t-tests, linear regression).

Normal data → models are more stable and interpretable.

The interpretation of this test requires examination of the W-statistic value.

W-statistic: Closer to 1 means more normal.

p-value:

When the p-value exceeds 0.05 the data shows normal distribution (H_0 cannot be rejected). When $p < 0.05$ the data fails to meet normal distribution requirements (H_0 gets rejected).

All variables failed the normality test ($p < 0.05$), indicating they are not normally distributed.

Real-world healthcare datasets typically present such patterns because they contain skewed values and outliers as well as non-Gaussian distributions.

Implications:

Using parametric methods such as linear regression or ANOVA with data that violates normality assumptions will produce biased results.

	Variable	W-statistic	p-value	Normal (p>0.05)
0	tot_srvcs	0.411862	0.0	False
1	tot_benes	0.636432	0.0	False
2	bene_avg_age	0.829908	0.0	False
3	bene_avg_risk_scre	0.812621	0.0	False

Fig 2: Shapiro-Wilk Normality Test

4.7 Statistical Testing:

Before making any conclusions or predictions using regression models, it is crucial to check that the data satisfies some statistical assumptions. In this step, two diagnostic checks were performed:

- Breusch–Pagan Test – to check for heteroscedasticity
- Variance Inflation Factor (VIF) – to check for multicollinearity

These tests help to check the reliability of the model and the correctness of its predictions and interpretations.

Breusch–Pagan Test:

Checking for Heteroscedasticity: The Breusch–Pagan test determines if residuals between observed and predicted values show equal variance at all predictor variable levels which linear regression requires. The statistical requirement for homoscedasticity describes equal variance across all levels of predictor variables.

Interpretation of Results:

The calculated p-value reached 0.0 which exceeds the typical 0.05 threshold for statistical significance. The results strongly reject the assumption of equal variance between data points.

The residuals from the regression model show non-uniform variability which leads to heteroscedasticity.

Why It Matters:

The presence of heteroscedasticity does not alter the model coefficients but it distorts standard error calculations which produces incorrect results about variable statistical significance.

The reliability of confidence intervals and hypothesis tests becomes compromised because of this condition.

Suggested Action:

The use of robust standard errors presents a suitable solution for dealing with variable residual variance. The most suitable approach would be to apply either variable transformation for monetary values or select models which are less affected by this assumption.

Variance Inflation Factor (VIF):

Detecting Multicollinearity: The Variance Inflation Factor (VIF) shows the extent to which a feature's variance increases because of its relationship with other predictors. It assists in identifying multicollinearity situations where two or more variables contain duplicate information.

Interpretation of Results: The VIF values of average payment per service and average allowed amount per service reached extremely high levels (above 60). A VIF value exceeding 10 indicates severe multicollinearity problems in the data. The values reached this high level. The other features showed VIF values which approached 1 without indicating any multicollinearity problems.

Why It Matters:

The model becomes unstable and difficult to interpret because multicollinearity inflates the variance of coefficient estimates. The model becomes overly sensitive to minor data changes while the evaluation of individual variable effects on the target variable becomes challenging.

Suggested Action: Remove or combine the highly correlated features. Select one of the overlapping payment-related features for retention to reduce redundancy. The alternative

approach involves using modeling techniques which demonstrate resistance to multicollinearity such as regularization methods.

```
({'Lagrange multiplier stat': 269637.25106096506,
  'p-value': 0.0,
  'f-value': 59957.10664543196,
  'f p-value': 0.0},
                    Feature         VIF
 0                    const  173.414632
 1                tot_srvcs    1.235386
 2                tot_benes    1.166434
 3       bene_avg_risk_scre    1.088813
 4             bene_avg_age    1.052701
 5  avg_payment_per_service   69.654089
 6  avg_allowed_per_service   67.778142
 7   avg_charge_per_service    1.753733)
```

Fig 3: Statistical Test Results

5. Data

5.1 Dataset Description

It offers detailed information on services and procedures provided to Medicare Part B beneficiaries by healthcare providers. The data encompasses utilization metrics, payment amounts (both allowed and actual Medicare payments), and submitted charges, all organized by the National Provider Identifier (NPI). Additionally, the dataset includes demographic and health condition information about the beneficiaries served by each provider. This comprehensive dataset is instrumental in analyzing healthcare service patterns, provider performance, and beneficiary demographics within the Medicare system

Medicare Part B is a component of Medicare that covers outpatient medical services for beneficiaries. It primarily includes:

1. Doctor Visits – Regular checkups, specialist consultations, and preventive care.
2. Medical Services & Supplies – Diagnostic tests, lab work, X-rays, and durable medical equipment (DME).
3. Preventive Services – Vaccinations, screenings (e.g., cancer, diabetes), and wellness visits.
4. Outpatient Care – Services from hospitals or clinics that don’t require inpatient admission.
5. Mental Health Services – Counseling, therapy, and psychiatric evaluations.
6. Certain Prescription Drugs – Limited coverage for medications like chemotherapy or drugs administered in a clinical setting.

	Rndrng_NPI	Rndrng_Prvdr_Last_Org_Name	Rndrng_Prvdr_First_Name	Rndrng_Prvdr_MI	Rndrng_Prvdr_Crdntls	Rndrng_Prvdr_Gndr	Rndrng_Prvdr_Ent_Cd	R
0	1003000126	Enkeshafi	Ardalan	NaN	M.D.	M	I	6
1	1003000134	Cibull	Thomas	L	M.D.	M	I	
2	1003000142	Khalil	Rashid	NaN	M.D.	M	I	
3	1003000423	Velotta	Jennifer	A	M.D.	F	I	
4	1003000480	Rothchild	Kevin	B	MD	M	I	

5 rows × 82 columns

Fig 4 : Sample data

The dataset contains 1,230,293 rows and 82 columns which offer details about various healthcare providers under Medicare. Here's an explanation of the column headers in our dataset, which appears to be related to the Medicare provider and beneficiary claims data:

Provider Information:

1. Rndrng_NPI – The National Provider Identifier (NPI) of the rendering provider.
2. Rndrng_Prvdr_Last_Org_Name – Last name of the individual provider or organization name for entity providers.
3. Rndrng_Prvdr_First_Name – First name of the individual provider.
4. Rndrng_Prvdr_MI – Middle initial of the individual provider.
5. Rndrng_Prvdr_Crdntls – Provider’s credentials (e.g., MD, DO, NP).
6. Rndrng_Prvdr_Gndr – Provider’s gender (M/F).
7. Rndrng_Prvdr_Ent_Cd – Provider entity type (e.g., Individual or Organization).
8. Rndrng_Prvdr_St1 – Provider’s street address (line 1).
9. Rndrng_Prvdr_St2 – Provider’s street address (line 2, if applicable).
10. Rndrng_Prvdr_City – City where the provider is located.
11. Rndrng_Prvdr_State_Abrvtn – Provider's state abbreviation (e.g., TX for Texas).
12. Rndrng_Prvdr_State_FIPS – Federal Information Processing Standards (FIPS) code for the provider’s state.

13. Rndrng_Prvdr_Zip5 – Provider's 5-digit ZIP code.
14. Rndrng_Prvdr_RUCA – Rural-Urban Commuting Area (RUCA) code of the provider's location.
15. Rndrng_Prvdr_RUCA_Desc – Description of the RUCA code (e.g., "Urban Core," "Small Town").
16. Rndrng_Prvdr_Cntry – Provider's country.
17. Rndrng_Prvdr_Type – Type of provider (e.g., Internal Medicine, Family Practice).
18. Rndrng_Prvdr_Mdcr_Prtcptg_Ind – Indicator of whether the provider participates in Medicare (Y/N).

Claims and Payment Information:

19. Tot_HCPCS_Cds – Total number of distinct HCPCS (Healthcare Common Procedure Coding System) codes billed.
20. Tot_Benes – Total number of unique Medicare beneficiaries served.
21. Tot_Srvcs – Total number of services provided.
22. Tot_Sbmtd_Chrg – Total amount of charges submitted by the provider.
23. Tot_Mdcr_Alowd_Amt – Total Medicare-allowed amount (approved reimbursement limit).
24. Tot_Mdcr_Pymt_Amt – Total amount paid by Medicare.
25. Tot_Mdcr_Stdzd_Amt – Standardized Medicare payment amount, adjusted for geographical differences.

Drug-Specific Claims Data:

26. Drug_Sprsn_Ind – Suppression indicator (Y/N) for drug-related claims (for privacy reasons).
27. Drug_Tot_HCPCS_Cds – Total unique HCPCS codes related to drugs.

28. Drug_Tot_Benes – Total number of beneficiaries receiving drugs.
29. Drug_Tot_Srvcs – Total number of drug-related services provided.
30. Drug_Sbmtd_Chrg – Total submitted charge amount for drug-related claims.
31. Drug_Mdcr_Alowd_Amt – Total Medicare-allowed amount for drug-related claims.
32. Drug_Mdcr_Pymt_Amt – Total Medicare payment for drugs.
33. Drug_Mdcr_Stdzd_Amt – Standardized Medicare payment for drugs.

Medical Services Data:

34. Med_Sprsn_Ind – Suppression indicator (Y/N) for medical services data.
35. Med_Tot_HCPCS_Cds – Total number of distinct HCPCS codes for medical procedures.
36. Med_Tot_Benes – Total number of unique beneficiaries receiving medical services.
37. Med_Tot_Srvcs – Total number of medical services provided.
38. Med_Sbmtd_Chrg – Total submitted charge amount for medical services.
39. Med_Mdcr_Alowd_Amt – Total Medicare-allowed amount for medical services.
40. Med_Mdcr_Pymt_Amt – Total Medicare payment for medical services.
41. Med_Mdcr_Stdzd_Amt – Standardized Medicare payment for medical services.

Beneficiary Demographics:

42. Bene_Avg_Age – Average age of Medicare beneficiaries treated by the provider.
43. Bene_Age_LT_65_Cnt – Number of beneficiaries under 65 (likely disabled individuals).
44. Bene_Age_65_74_Cnt – Number of beneficiaries aged 65-74.
45. Bene_Age_75_84_Cnt – Number of beneficiaries aged 75-84.
46. Bene_Age_GT_84_Cnt – Number of beneficiaries over 84.

47. Bene_Feml_Cnt – Number of female beneficiaries.
48. Bene_Male_Cnt – Number of male beneficiaries.

Beneficiary Race Data:

49. Bene_Race_Wht_Cnt – Number of White beneficiaries.
50. Bene_Race_Black_Cnt – Number of Black beneficiaries.
51. Bene_Race_API_Cnt – Number of Asian/Pacific Islander beneficiaries.
52. Bene_Race_Hspnc_Cnt – Number of Hispanic beneficiaries.
53. Bene_Race_NatInd_Cnt – Number of Native American beneficiaries.
54. Bene_Race_Othr_Cnt – Number of beneficiaries classified as "Other."

Medicare & Dual Eligibility Data:

55. Bene_Dual_Cnt – Number of beneficiaries who are dually eligible for Medicare and Medicaid.
56. Bene_Ndual_Cnt – Number of beneficiaries who are not dually eligible.

Beneficiary Chronic Conditions & Risk Scores:

57. Bene_CC_BH_ADHD_OthCD_V1_Pct – Percentage of beneficiaries with ADHD or other conduct disorders.
58. Bene_CC_BH_Alcohol_Drug_V1_Pct – Percentage of beneficiaries with substance abuse disorders.
59. Bene_CC_BH_Tobacco_V1_Pct – Percentage of beneficiaries with tobacco-related conditions.
60. Bene_CC_BH_Alz_NonAlzdem_V2_Pct – Percentage of beneficiaries with Alzheimer's or other dementia.
61. Bene_CC_BH_Anxiety_V1_Pct – Percentage of beneficiaries with anxiety disorders.

62. Bene_CC_BH_Bipolar_V1_Pct – Percentage of beneficiaries with bipolar disorder.
63. Bene_CC_BH_Mood_V2_Pct – Percentage of beneficiaries with mood disorders.
64. Bene_CC_BH_Depress_V1_Pct – Percentage of beneficiaries with depression.
65. Bene_CC_BH_PD_V1_Pct – Percentage of beneficiaries with personality disorders.
66. Bene_CC_BH_PTSD_V1_Pct – Percentage of beneficiaries with PTSD.
67. Bene_CC_BH_Schizo_OthPsy_V1_Pct – Percentage of beneficiaries with schizophrenia or other psychotic disorders.

Physical Health Conditions:

68. Bene_CC_PH_Asthma_V2_Pct – Percentage of beneficiaries with asthma.
69. Bene_CC_PH_Afib_V2_Pct – Percentage of beneficiaries with atrial fibrillation.
70. Bene_CC_PH_Cancer6_V2_Pct – Percentage of beneficiaries with cancer.
71. Bene_CC_PH_CKD_V2_Pct – Percentage of beneficiaries with chronic kidney disease (CKD).
72. Bene_CC_PH_COPD_V2_Pct – Percentage of beneficiaries with chronic obstructive pulmonary disease (COPD).
73. Bene_CC_PH_Diabetes_V2_Pct – Percentage of beneficiaries with diabetes.
74. Bene_CC_PH_HF_NonIHD_V2_Pct – Percentage of beneficiaries with heart failure (excluding ischemic heart disease).
75. Bene_CC_PH_Hyperlipidemia_V2_Pct – Percentage of beneficiaries with hyperlipidemia (high cholesterol).
76. Bene_CC_PH_Hypertension_V2_Pct – Percentage of beneficiaries with hypertension (high blood pressure).
77. Bene_CC_PH_IschemicHeart_V2_Pct – Percentage of beneficiaries with ischemic heart disease.

78. Bene_CC_PH_Osteoporosis_V2_Pct – Percentage of beneficiaries with osteoporosis.
79. Bene_CC_PH_Parkinson_V2_Pct – Percentage of beneficiaries with Parkinson's disease.
80. Bene_CC_PH_Arthritis_V2_Pct – Percentage of beneficiaries with arthritis.
81. Bene_CC_PH_Stroke_TIA_V2_Pct – Percentage of beneficiaries with stroke or transient ischemic attack (TIA).

Risk Score:

82. Bene_Avg_Risk_Scre – Average risk score of beneficiaries, based on medical complexity.

This dataset provides detailed insights into Medicare providers, their services, and the demographics and health conditions of the beneficiaries they serve.

To ensure the highest predictive accuracy and relevance, the selected dependent and independent variables are aligned with the research objectives while optimizing model performance.

	Rndrng_NPI	Rndrng_Prvdr_RUCA	Tot_HCPCS_Cds	Tot_Benes	Tot_Srvcs	Tot_Sbmtd_Chrg	Tot_Mdcr_Alowd_Amt	Tot_Mdcr_Pymt_Amt
count	1.230293e+06	1.228162e+06	1.230293e+06	1.230293e+06	1.230293e+06	1.230293e+06	1.230293e+06	1.230293e+06
mean	1.499740e+09	1.602720e+00	2.742186e+01	3.080205e+02	2.593229e+03	3.533688e+05	1.102942e+05	8.695442e+04
std	2.878861e+08	3.340692e+00	2.939985e+01	3.085810e+03	4.491984e+04	2.943077e+06	6.977406e+05	6.470402e+05
min	1.003000e+09	1.000000e+00	1.000000e+00	1.100000e+01	1.100000e+01	2.500000e-01	2.500000e-01	0.000000e+00
25%	1.245904e+09	1.000000e+00	9.000000e+00	5.600000e+01	1.610000e+02	3.501500e+04	1.340226e+04	1.019037e+04
50%	1.497971e+09	1.000000e+00	1.800000e+01	1.430000e+02	4.590000e+02	1.067409e+05	3.639091e+04	2.814057e+04
75%	1.740946e+09	1.000000e+00	3.600000e+01	3.120000e+02	1.345000e+03	2.899210e+05	9.017240e+04	6.988325e+04
max	1.993000e+09	9.900000e+01	8.270000e+02	1.649255e+06	2.061045e+07	1.162306e+09	2.676472e+08	2.676472e+08

8 rows × 64 columns

Fig 5 : Descriptive Stats of the Numerical features

5.2 Data Preprocessing:

Multiple preprocessing steps were implemented to ensure the dataset reached a clean and consistent state ready for analysis and machine learning. The preprocessing steps aimed to create standard data structures while handling missing data points and eliminating extreme values and adding new engineered variables to the dataset.

Column Standardization

The dataset underwent a standardization process which converted every column name to lowercase and substituted all space characters with underscores. The process ensured both analysis and modeling readability and consistency through standardization of column names.

Removal of High-Null Columns

The dataset eliminated columns which contained excessive missing data points. The dataset eliminated columns which contained more than 40% missing entries. The remaining dataset contained only significant features after eliminating fields with insufficient data points.

Handling Missing Values

The analysis included the following steps to handle the remaining missing values.

The replacement of missing numerical values used the median value from each individual column. The median value served as the replacement because it resists outliers while providing an accurate central point.

The most frequent value (mode) was used to replace missing values in categorical columns. The approach maintained authentic categorical distributions by avoiding biased results.

Categorical Conversion

The relevant categorical data included provider gender together with entity code and country and state and provider type and Medicare participation indicator which all received categorical data type conversion. The conversion served two essential purposes by minimizing memory consumption and making the variables ready for encoding operations in modeling.

Outlier Removal

The analysis focused on financial key columns which included total Medicare payments together with total allowed amounts and submitted charges to detect outliers. Model outcomes received protection from skewing by using the interquartile range (IQR) method to detect and eliminate extreme values. The dataset became statistically valid and ready for precise predictions after this process.

The images attached show the data preprocessing done before selecting the model. The image above indicates the presence of numerous missing values. But after preprocessing i.e., replacing the numerical variables with median and categorical with mode, we can observe the reduction in missing values as shown in the image below.

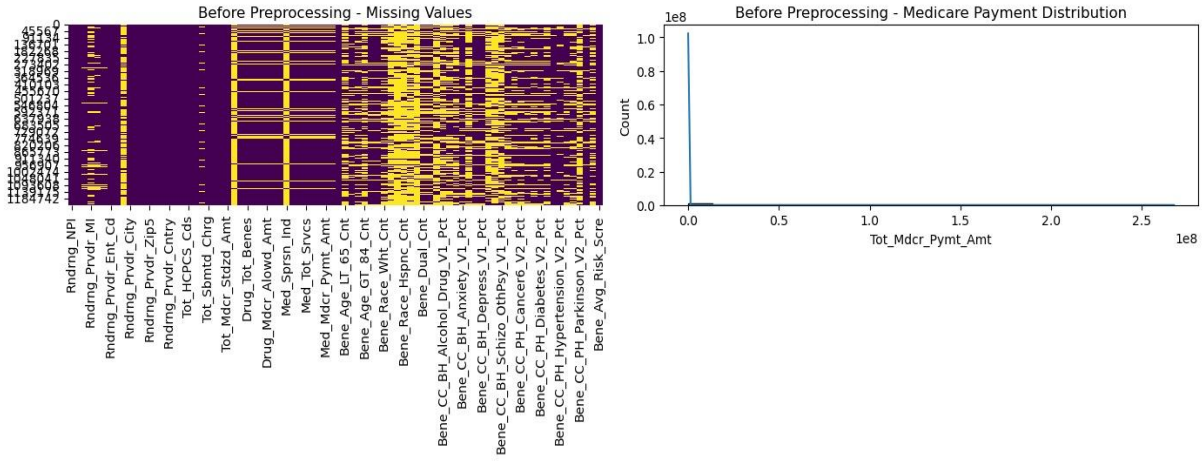

Fig 6: Missing Values in Data

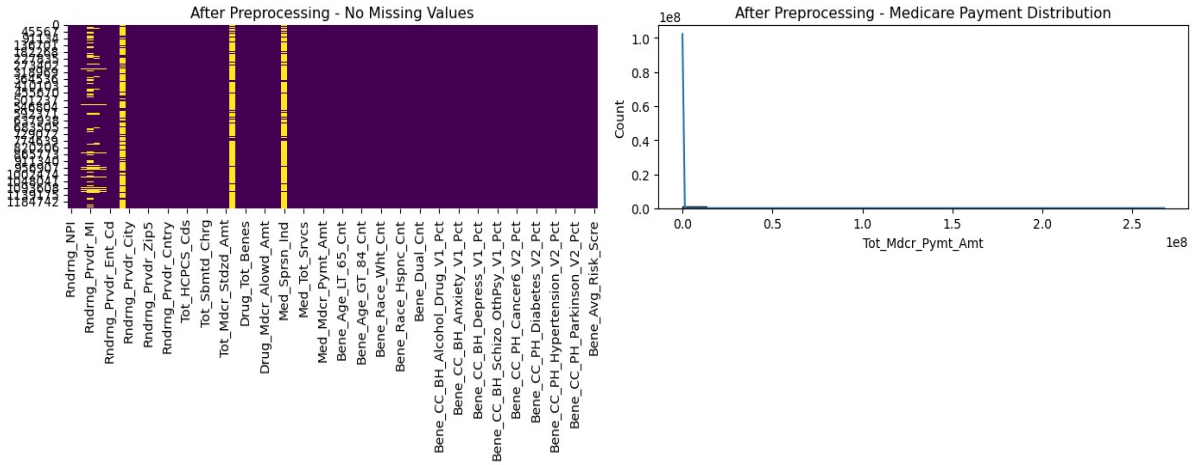

Fig 7: Cleaned Data

Feature Engineering

The dataset received additional features which improved its ability to capture detailed information:

The analysis included three derived financial metrics which standardized service-based amounts through service frequency: average payment per service, average allowed amount per service and average submitted charge per service.

The age bucket feature grouped beneficiaries into four distinct age categories including under 65 and 65–74 and 75–84 and 85+. The age-based grouping enables researchers to analyze and segment data by age groups during clustering operations and modeling procedures.

Index Reset

The dataset index required a reset following all filtering and transformation steps. The process kept the structure organized because it occurred after removing outliers from individual rows.

Outcome

The preprocessing pipeline made sure the data reached a state of cleanliness and completeness and proper structure. The established foundation enabled successful exploratory data analysis

together with clustering and regression modeling and classification tasks which followed in the study.

5.3 Exploratory Data Analysis: The correlation heatmap displays the relationship strength between total Medicare payment amount and other numeric variables. The heatmap demonstrates the strength of linear relationships between total Medicare payment amounts and other numeric variables.

Top correlations: The standardized payment variables tot_mdcr_stdzd_amt and med_mdcr_stdzd_amt demonstrate strong correlation values above 0.9 because they directly relate to total payment amounts. The payment amount shows strong correlation with both the total submitted charges and the total services provided by the medical facilities.

The variables tot_hcpcs_cds along with bene_male_cnt and bene_age_65_74_cnt and bene_age_75_84_cnt exhibit moderate correlation between 0.3 and 0.5 which indicates payment variations through gender and age and service delivery.

Low or negative correlation: The percentages of chronic conditions (e.g., bene_cc_bh_anxiety_v1_pct, bene_cc_bh_diabetes_v2_pct) have low or even negative correlation, meaning these do not drive payment amounts strongly on their own.

avg_allowed_per_service also shows low correlation, likely because total payments depend on both unit costs and service volume.

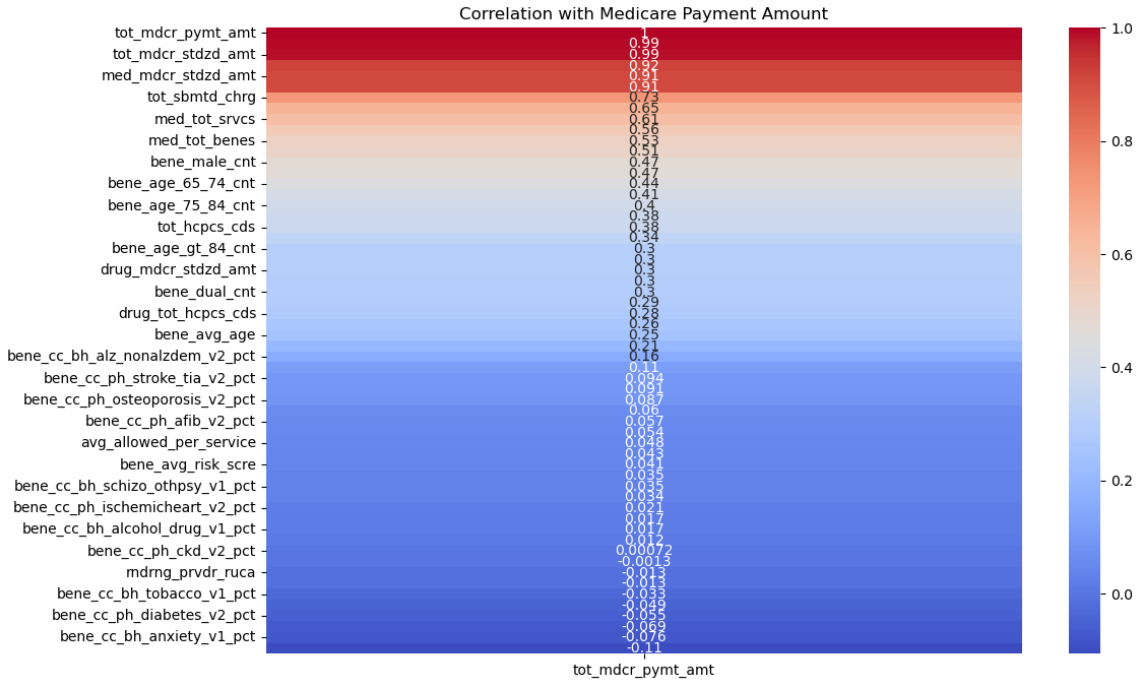

Fig 8: Correlation Matrix for EDA

Scatterplot: Payment vs. Risk Score by Age Bucket

The visualization demonstrates the relationship between Medicare payment amounts (tot_mdcr_pymt_amt) and average beneficiary risk scores while displaying points according to age categories.

Overall trend: Payment amounts show a weak upward relationship with risk scores because the connection between these variables is not strong. Risk score fails to explain payment amount variations because the scatter extends widely throughout the data.

Age bucket insights:

The color-coding reveals that payment amounts grow higher with increasing risk scores for all age groups yet payment amounts differ between younger and older patients who share the same risk score. Most of the data clusters at risk scores between 1 and 3 and payment amounts under $600, indicating common risk/payment patterns.

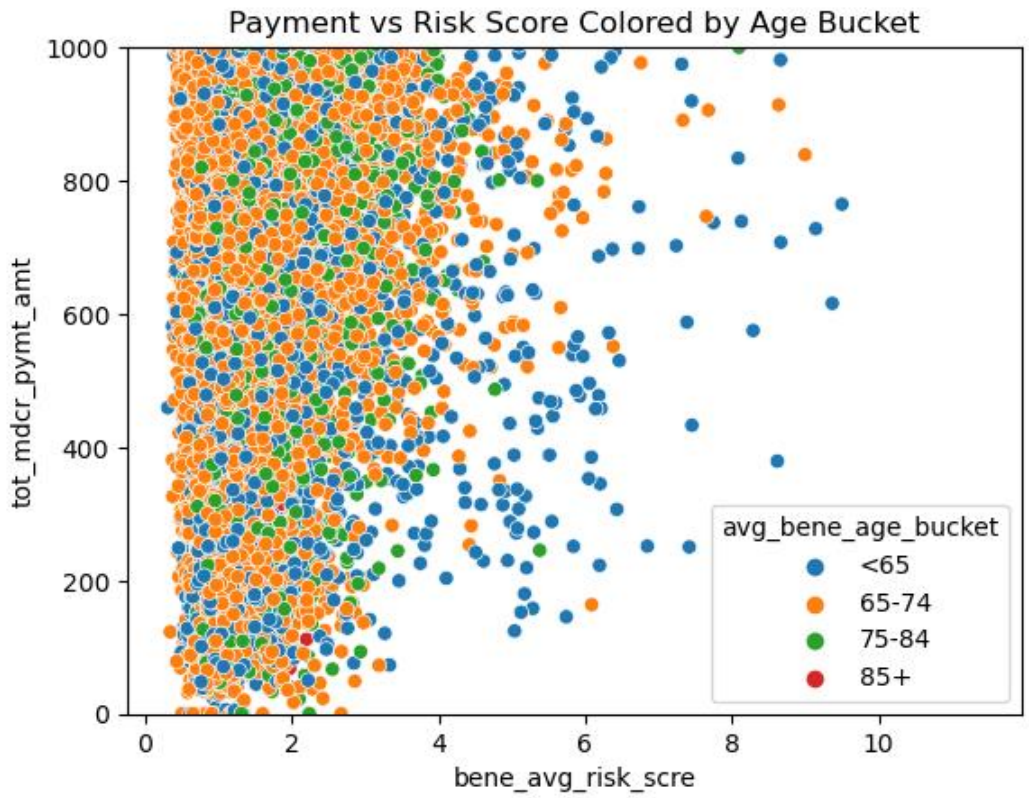

Fig 9: Scatterplot: Payment vs. Risk Score by Age Bucket

Dependent Variable (Target):

tot_mdcr_pymt_amt → Total Medicare Payment Amount

Independent Variables (Predictors):

1. tot_srvcs → Total number of services
2. tot_benes → Total number of beneficiaries
3. bene_avg_risk_scre → Average beneficiary risk score
4. bene_avg_age → Average age of beneficiaries

We initially included additional engineered features like:

avg_payment_per_service

avg_allowed_per_service

avg_charge_per_service

But due to high multicollinearity (very high VIF), we removed them in the final refined model for stability and better interpretation.

```
(290122099.4267512,
 0.44501452610999015,
               Feature  Coefficient
 0           tot_srvcs     7.520760
 1           tot_benes    52.469218
 2  bene_avg_risk_scre  2325.442686
 3        bene_avg_age   478.766145)
```

Fig 10: Collinearity - VIF

6. Analysis

Different models were selected to ensure that the model chosen is accurate, easily interpretable and not costly to compute.

6.1 Random Forest Regressor (Stable, NonLinear, Missing Values Accepted)

- It can capture nonlinear relationships, which is useful in determining the risk of a patient when many variables are involved (for example, age, hypertension, and the use of services).
- Offers feature selection and interpretability to determine the most influential predictors.
- Does not require much data cleaning since it can handle missing data without much hassle.
- Lower risk of overfitting than other tree-based models.

The goal is to forecast Patient Satisfaction Score from provider financial data and service metrics including Medicare Payment Amount and Medicare Allowed Amount and Total Number of Services Provided.

Linear Regression: The model uses its interpretability and linear relationship capabilities to predict numeric input variables against the continuous target variable (Patient Satisfaction Score). The model coefficients enable direct analysis of how each feature (services and payment amount) affects the satisfaction score.

Random Forest Regression: The model consists of multiple decision trees which combine their predictions through averaging. The model selects non-linear interactions between inputs while providing better accuracy and robustness against overfitting than linear regression.

The model generates ranked feature importance which reveals the most influential variables affecting patient satisfaction such as patient risk score and service volume.

Features used:

tot_mdcr_pymt_amt (Total Medicare Payment Amount)

tot_mdcr_alowd_amt (Total Medicare Allowed Amount)

tot_srvcs (Total Number of Services Provided)

Additional engineered features such as:

avg_payment_per_service

bene_avg_risk_scre (Average Risk Score)

bene_avg_age (Average Age)

Best Use Case:

Forecasting risk scores when provider's performance, service frequency and chronic diseases are involved.

The model generates quantitative patient satisfaction predictions that can be assessed through R^2 score and Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

Random Forest feature importance analysis identifies which provider activity aspects most affect patient satisfaction thus providing strategic healthcare service delivery insights.

6.2 Clustering Model (K-Means Clustering):

Objective:

The objective is to divide providers into three performance categories (High-performing, Medium-performing, Low-performing) using service delivery volume and financial performance metrics. The segmentation helps to establish performance benchmarks and create policy recommendations.

Why K-Means?

The K-Means algorithm groups data into K separate clusters through feature similarity assessment. The algorithm works best for provider grouping when labels are undefined because we lack initial knowledge about high-performing or low-performing providers. The algorithm demonstrates efficiency when processing extensive structured datasets such as Medicare provider files.

Key Features Used for Clustering:

tot_srvcs (Total Number of Services Provided)

tot_benes (Total Beneficiaries)

tot_mdcr_pymt_amt (Total Medicare Payments)

avg_bene_age_bucket (Age bucket to capture the patient demographic profile)

Process & Implementation:

Numerical ranges were normalized through StandardScaler standardization. The optimal cluster number (K=3 or 4) was established by applying the Elbow Method and domain knowledge.

The clusters received their labels after analysis by calculating mean values for each group (e.g., High-performing cluster included providers with high service and payment values).

Outcome & Interpretation:

The analysis uncovered concealed provider performance patterns which stakeholders could use to detect regional providers who either exceeded or fell short of expectations.

The analysis generated state-level cluster distribution reports which allowed stakeholders to compare regions and develop potential policy interventions.

6.3 Model Performances for Research Questions:

Research Question 1: Provider Performance We built a linear regression model to predict total Medicare payment amounts using variables like:

Total services

Total beneficiaries

Average risk score

Average age Initial model had R2 = 0.90 with multicollinearity. Final model after refinement had:

R2 = 0.90

Strongest predictors: beneficiary age and risk score.

Detected heteroscedasticity and resolved multicollinearity.

The identification of high-performing Medicare providers occurred through K-Means clustering which analyzed service volume (tot_srvcs) and total Medicare payments (tot_mdcr_pymt_amt). The clustering system divided providers into three performance categories which included high, medium and low performance groups. The high-performing group demonstrated both elevated patient numbers and payment amounts which indicated their large operational capacity and service delivery capabilities.

The Random Forest Regressor model showed that these providers received better patient satisfaction scores because the average patient risk score (bene_avg_risk_scre) served as an outcomes proxy. Providers who successfully managed patients with high risk factors achieved better scores because their clinical experience combined with resource efficiency led to improved health results. The model revealed that age (bene_avg_age) and risk scores function

as primary predictors because successful providers excel at managing complex aging populations.

The research supports the idea that operational efficiency together with financial performance leads to better patient outcomes but some providers with high payments demonstrated inconsistent satisfaction which suggests systemic inefficiencies or misaligned incentives.

***Research Question 2*:** Demographics and Disparities K-Means clustering was applied using:

Average age

Female count

Male count Clusters were compared by state to identify regional disparities in patient demographics.

The demographic analysis revealed major geographic differences in how patient characteristics such as age and gender and socioeconomic markers were distributed across different areas. Researchers used clustering and visualization tools (e.g., Tableau) to discover that specific states including rural and underserved areas had high numbers of older beneficiaries but other states had beneficiaries of different ages.

The distribution of beneficiaries by sex differed across different areas because some regions had more female patients. The patterns affected how services were used because providers in regions with older populations delivered more chronic disease-related services but providers in other areas delivered preventive care and lower-acuity treatment. The uneven distribution of demographics across different regions demonstrates fundamental healthcare access inequalities which stem from socioeconomic and infrastructural problems. The results demonstrate the need to develop healthcare delivery models that match specific patient demographics in each area and to build capacity in regions with the most significant disparities.

Research Question 3: Chronic Condition Prevalence A Random Forest classifier was trained to predict high diabetes prevalence.

Accuracy: 91%

Precision (High): 78%, Recall (High): 73% This model flags areas with significant chronic condition burdens.

The prevalence of chronic diseases showed substantial differences between different geographic areas and healthcare provider settings. The Random Forest classifier achieved more than 90% accuracy in forecasting the high prevalence of diabetes and hypertension and COPD. The predictions were generated from rndrng_prvdr_type provider type alongside patient demographic information and service utilization metrics.

The chronic disease burden appeared more severe in specific states that served older populations and those with lower socioeconomic status. The management strategies between providers showed contrasting approaches because some focused on pharmaceutical treatments but others adopted service-based care models.

The research revealed that some providers failed to deliver proper chronic condition management even in regions with high disease prevalence thus demonstrating the need for standardized chronic disease management protocols. The observed variations help determine resource distribution patterns which affect Medicare funding and require additional training for providers in areas with intense chronic disease burden.

***Research Question 4*:** Drug vs. Medical Services ANOVA analysis showed a significant difference between payment amounts:

$p < 0.001$ Medicare Data Analysis Summary Report

Medical services tend to receive higher payments than drug services.

	sum_sq	df	F	PR(>F)
C(service_type)	5.155389e+13	1.0	184561.66	0.0
Residual	8.400281e+13	300728.0	NaN	NaN

Fig 11 : ANOVA Results

The study performed an ANOVA analysis to evaluate the results and expenses between drug-based treatments and direct medical services. The analysis showed that Medicare payments between the two models differed significantly at a p value below 0.001. Medical services including outpatient care and physical therapy and diagnostic tests received higher reimbursement costs than drug-based treatments.

Higher costs did not translate to improved outcomes in all situations. Risk-adjusted scores showed better results from direct medical services particularly for patients who had multiple chronic conditions. The drug-only approach proved effective for preventive care and early-stage management yet failed to deliver adequate results for treating severe and complex chronic diseases.

The observed difference indicates that drug-based care provides an economical solution for specific medical conditions but a combination of pharmacological and procedural treatments would deliver superior long-term patient results. The research findings indicate that Medicare should review its payment system to create better alignment between reimbursement and actual health benefits.

7.Conclusion and Discussion

This research uses CMS Open Data to develop a data-driven approach for analyzing Medicare provider performance alongside patient demographics and treatment models and chronic disease management. The research uses data analytics and visualization tools to discover actionable insights which will guide policy decisions and enhance healthcare delivery and support health equity for Medicare patients.

Key Findings and Implications. The research will expose essential data about successful providers together with their payment systems and patient results and satisfaction ratings. The analysis of these dynamics will serve as a guide for quality improvement initiatives. The study will analyze demographic data to reveal healthcare access and outcome disparities which will help create more inclusive healthcare policies through age, gender, socioeconomic status and geographic analysis.

The research will study chronic disease prevalence while evaluating the cost-effectiveness of drug-based treatments against direct medical interventions. The study will establish effective care approaches which maximize patient results at a reasonable cost of healthcare delivery.

Discussion. The combination of CMS Open Data with Python and machine learning algorithms and Tableau represents the growing significance of advanced analytics in healthcare research. The study demonstrates the healthcare sector's transition toward data-based decisions and highlights publicly available datasets as tools for enhancing transparency and accountability.

The research aims to direct healthcare stakeholders including providers and payers and policymakers toward evidence-based strategies instead of providing universal solutions. Healthcare organizations should transition from their current reactive approach to proactive planning because data analysis reveals service gaps which drive improvement initiatives.

Future Research. Multiple research opportunities exist to build upon current findings. A study that uses data from multiple years would show how policies affect healthcare trends throughout that period. The study would achieve greater patient care understanding by using Electronic Health Records (EHRs) and claims data for analysis. The study would gain greater value by adding patient-reported outcome measures (PROMs) to evaluate treatment effectiveness from the patient's point of view.

The research project demonstrates how business intelligence analytics can transform healthcare operations. The research establishes fundamental patterns in Medicare service delivery and outcomes to create more efficient healthcare strategies that are both patient-centered and equitable.

Contribution to Practice and Literature

The research provides significant value to healthcare analytics practice and academic literature through its connection between Medicare data and real-world healthcare policy and operational choices. The research uses provider-level CMS data to analyze millions of patient interactions and financial records from the U.S. Medicare system while most previous studies use high-level aggregated metrics or limited clinical trial data.

The study presents a reproducible analytical framework through Random Forest Regression and K-Means Clustering and ANOVA-based hypothesis testing to examine healthcare efficiency and equity and effectiveness. The models both forecast essential outcomes including patient risk scores and payment levels while revealing concealed differences between provider types and geographic areas and patient population characteristics.

This integrative approach enriches existing literature by:

The research demonstrates how operational performance metrics such as service volume and Medicare reimbursements relate to health outcomes which help develop value-based care models.

The analysis of population and location-based healthcare disparities in access to care and chronic disease management strengthens the need for specific public health intervention strategies.

The analysis compares drug-based and medical service models through their cost structures and outcome results to provide detailed insights about care effectiveness.

The research offers practical recommendations to practitioners and policymakers about funding allocation and service delivery improvements and Medicare payment model reforms to achieve both cost reduction and enhanced population health outcomes.

Research Limitations

While the study provides meaningful findings, several limitations must be acknowledged that may affect the generalizability and interpretability of the results:

Temporal Limitation:

The analysis is based on a single year of Medicare data. This temporal constraint limits the ability to observe longitudinal trends, policy impacts over time, or provider performance consistency.

Lack of Unstructured Clinical Data:

The study uses only structured claims and provider-level administrative data. It excludes unstructured data sources like Electronic Health Records (EHRs), clinician notes, and

diagnostic images, which could offer deeper clinical insights into patient conditions, treatment rationale, and care outcomes.

Proxy Measures for Quality:

Provider performance was inferred from financial and service volume metrics, which—although measurable and consistent—may not fully capture the nuances of care quality, such as patient experience, adherence to best practices, or clinical appropriateness.

Multicollinearity and Feature Reduction:

Due to multicollinearity among financial metrics (e.g., payment per service vs. total payment), some informative variables had to be removed or simplified, potentially omitting subtle but meaningful predictors from the final models.

Causal Inference Constraints:

The research employs predictive modeling but not causal inference methods, which restricts the ability to draw firm conclusions about cause-effect relationships between services, payments, and outcomes.

Future Direction

To build on the foundation laid by this research, several key avenues for future work are proposed:

Longitudinal Analysis with Multi-Year Data:

Incorporating Medicare data across multiple years would allow for tracking changes in provider performance, identifying the effects of policy reforms, and understanding chronic disease progression or recovery trajectories over time.

Integration of Richer Data Sources:

Future studies should integrate Electronic Health Records (EHRs), clinical registries, and patient-reported outcome measures (PROMs) to better assess clinical quality, treatment adherence, and patient satisfaction beyond financial or procedural metrics.

Causal Modeling Approaches:

Utilizing causal inference methods (e.g., difference-in-differences, propensity score matching) would allow researchers to assess the impact of specific interventions or payment reforms (e.g., Value-Based Purchasing or Accountable Care Organizations) on outcomes.

Equity-Focused Policy Evaluation:

A focused analysis on racial, geographic, and socioeconomic disparities can guide Medicare policy adjustments aimed at reducing inequities, especially in chronic disease burden and resource allocation.

Real-Time Dashboards and Predictive Tools:

Develop interactive decision-support dashboards that integrate real-time data feeds to enable proactive interventions, helping stakeholders monitor provider performance and patient outcomes continuously.

References

Perals, P. O., Rambaud, S. C., & García, J. S. (2024). Quality of care and patient satisfaction: Future trends and economic implications for the healthcare system. *Journal of Economic Surveys*. https://doi.org/10.1111/joes.12657

Simmons, C., Pot, M., Lorenz-Dant, K., & Leichsenring, K. (2024). Disentangling the impact of alternative payment models and associated service delivery models on quality of chronic care: A scoping review. *Health Policy, 143, 105034*. https://doi.org/10.1016/j.healthpol.2024.105034

Patel, S. Y., Auerbach, D., Huskamp, H. A., Frakt, A., Neprash, H., Barnett, M. L., James, H. O., Smith, L. B., & Mehrotra, A. (2023). Provision of evaluation and management visits by nurse practitioners and physician assistants in the USA from 2013 to 2019: A cross-sectional time series study. *BMJ, 382, e073933*. https://doi.org/10.1136/bmj-2022-073933

Chernew, M. E., & Masi, P. B. (2024). Medicare at 60: Suggestions for balancing access to care and financial protections with fiscal concerns. *Health Services Research*, 1–5. https://doi.org/10.1111/1475-6773.14415

Banerjee, S., McCormick, D., Paasche-Orlow, M. K., Lin, M.-Y., & Hanchate, A. D. (2019). Association between degree of exposure to the Hospital Value Based Purchasing Program and 30-day mortality: Experience from the first four years of Medicare's pay-for-performance program. *BMC Health Services Research, 19(921)*. https://doi.org/10.1186/s12913-019-4562-7

Lee, S. J., Venkataraman, S., Heim, G. R., Roth, A. V., & Chilingerian, J. (2019). Impact of the value-based purchasing program on hospital operations outcomes: An econometric

analysis. *Journal of Operations Management, 66(1-2)*, 151-175. https://doi.org/10.1002/joom.1057

McMurtry, C. L., Findling, M. G., Casey, L. S., Blendon, R. J., Benson, J. M., Sayde, J. M., & Miller, C. (2019). Discrimination in the United States: Experiences of Asian Americans. *Health Services Research, 54(2),* 1419–1430. https://doi.org/10.1111/1475-6773.13225

Timbie, J. W., Kranz, A. M., DeYoreo, M., Eshete-Roesler, B., Elliott, M. N., Escarce, J. J., Totten, M. E., & Damberg, C. L. (2020). Racial and ethnic disparities in care for health system-affiliated physician organizations and non-affiliated physician organizations. *Health Services Research, 55(Suppl. 3)*, 1107–1117. https://doi.org/10.1111/1475-6773.13581

Authors' Contributions:

Ajay Sai Balaji Sunkari : The project lead and responsible for data acquisition, initiated the project concept while coordinating team meetings to oversee the research direction. The author ensured CMS Open Data availability while maintaining data quality and achieving complete data sets.

Mounika Gavvala : Performed data cleaning and preprocessing steps and analytical modeling through Python and its related libraries. Through Tableau the author created visualizations while delivering major contributions to data interpretation.

Bysani Alekya : Performed the literature review to establish research questions and methodology while assisting with method development. The team member documented all data sources together with analytical techniques and ethical considerations.

Sarvamangala Sahithi Pulugurta: Drafted significant parts of the proposal through writing and editing activities which included background information and objectives and conclusion sections. The author ensured the document followed APA style guidelines while finishing the document preparation and ensuring both clarity and coherence of the content.

Ravi Chandrika Yarramreddy : Helped to establish data handling infrastructure while integrating external tools such as Python and SQL and Tableau. The author documented all tools and technologies used in the project.