An Explainable Machine Learning Framework for Predicting Healthcare Utilization and Quantifying Economic Burden in US Health Systems

Samiha Binte Abdullah., Anurodh Singh., Gaurav Kudeshia

Master of Science in Business Analytics Ambassador Crawford College of Business and Entrepreneurship Kent State University, Kent, Ohio, USA

DOI: https://doi.org/10.51244/IJRSI.2026.1306000345

Received: 17 June 2026; Accepted: 22 June 2026; Published: 10 July 2026

ABSTRACT

Background: Rising healthcare expenditure in the United States represents one of the most critical challenges facing health systems today. Accurate prediction of healthcare utilisation patterns and associated economic burden is essential for equitable resource allocation, insurance planning, and evidence-based health policy.

Objective: This study presents an explainable machine learning framework integrating a two-stage economic modelling architecture with Shapley Additive Explanations (SHAP) to simultaneously predict healthcare utilisation across three care settings and quantify individual and system-level economic burden.

Methods: We utilise the Medical Expenditure Panel Survey (MEPS) as the primary dataset, supplemented by HCUP for external validation. The pipeline encompasses structured preprocessing, novel composite feature engineering, and competitive benchmarking of XGBoost, Random Forest, LightGBM, and baseline linear regression. A two-stage economic model first predicts utilisation, then generates cost estimates conditional on those predictions.

Results: XGBoost achieved superior performance (RMSE = 1.24 ED visits, R2 = 0.847 inpatient admissions) with an MAE of $1,847 per patient for total expenditure. SHAP decomposition identified chronic disease burden, age, insurance coverage depth, prior utilisation, and socioeconomic vulnerability as the five primary cost drivers. System-level forecasts matched CMS figures within 3.2%.

Conclusions: The framework advances the state of the art by unifying utilisation prediction, econometric modelling, and explainable AI in a single reproducible pipeline. Its transparency makes it directly applicable to health policy resource allocation, insurance planning, and algorithmic accountability.

Keywords: Machine learning; Healthcare utilisation; Economic burden; SHAP explainability; XGBoost; Medical expenditure; Predictive modelling; Health equity; United States health systems

INTRODUCTION

Background and Motivation

The United States healthcare system accounts for approximately 17.3% of gross domestic product, with aggregate national health expenditure exceeding $4.5 trillion annually. This figure disproportionately large relative to comparable high-income nations reflects deep structural inefficiencies in service delivery, insurance market fragmentation, and the chronic under-management of high-utilisation patient cohorts. Emergency department overcrowding, preventable hospital readmissions, and catastrophic out-of-pocket expenditure represent the most visible manifestations of these systemic failures.

Accurate predictive modelling of healthcare utilisation and associated costs offers a powerful lever for systems-level reform. If health administrators, insurers, and policymakers can reliably anticipate which patient cohorts

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

will generate disproportionate demand and at what projected cost they can proactively deploy targeted interventions. Yet existing predictive tools suffer from two interrelated limitations: insufficient predictive accuracy due to neglect of non-linear interaction effects, and opacity that prevents clinical and regulatory acceptance.

The emergence of gradient boosting architecture, particularly XGBoost and LightGBM alongside post-hoc explainability frameworks such as SHAP has created an opportunity to resolve both limitations simultaneously. This paper capitalises on that opportunity by presenting a fully integrated, reproducible pipeline combining state-of-the-art ML prediction with a novel two-stage economic model and comprehensive SHAP-based transparency.

Problem Statement

Three discrete gaps motivate this research:

1. Fragmented prediction pipelines: existing studies separately address utilisation prediction or cost estimation, but rarely integrate both within a single validated framework.

2. Absence of two-stage economic decomposition: cost is frequently modelled as a direct function of demographics, ignoring the causal pathway through clinical demand. This produces biased cost estimates and limits policy actionability.

3. Interpretability deficit: black-box models preclude adoption in clinical governance and regulatory contexts, necessitating tools that demonstrate algorithmic accountability.

Research Objectives

This study pursues four primary objectives:

4. To construct a reproducible ML pipeline for predicting ED visits, inpatient admissions, and outpatient encounters using MEPS data.

5. To develop and validate a two-stage economic model conditioning cost estimation on predicted utilisation outputs.

6. To apply SHAP explainability methods to decompose feature-level drivers of utilisation and cost at both individual and population levels.

7. To generate system-level expenditure forecasts and risk stratification profiles applicable to health policy and insurance planning.

Novelty and Contributions

- An integrated utilisation-cost prediction framework bridging clinical demand forecasting and health economic modelling.

- A two-stage economic decomposition model separating clinical demand from financial impact.

- A multi-dimensional SHAP analysis producing feature importance rankings, individual prediction explanations, and equity bias assessments.

- A fully reproducible open-source pipeline enabling replication and extension by health data scientists and policy researchers.

LITERATURE REVIEW

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

Machine Learning in Healthcare Utilisation Prediction

The application of supervised ML to healthcare utilisation prediction has grown substantially over the past decade. Within the utilisation domain, random forest models have been applied to predict 30-day hospital readmissions (Futoma et al., 2015), while gradient boosting approaches have shown superior performance in predicting ED revisits in paediatric populations (Goto et al., 2019). However, the majority of utilisation prediction studies address a single care setting in isolation. Cross-setting integrated models capable of simultaneously forecasting ED visits, inpatient admissions, and outpatient encounters remain rare in the literature.

Economic Burden Modelling in US Healthcare

The economic modelling literature has primarily relied on generalised linear models, two-part models, and OLS regression with log-transformed outcomes. Manning and Mullahy (2001) established the canonical econometric framework for healthcare cost modelling. ML-based cost prediction models have demonstrated significant improvements over GLM baselines when non-linear predictor interactions are present. Yet few studies have explicitly modelled the pathway from utilization to cost treating cost as a direct outcome of patient characteristics rather than as a function of utilised services.

Explainable AI in Clinical Decision Support

The explainability of ML models in healthcare has emerged as a critical requirement for ethical deployment and regulatory compliance. SHAP introduced by Lundberg and Lee (2017) has become the dominant post-hoc explainability method due to its theoretical grounding in cooperative game theory. Recent studies have applied SHAP to sepsis prediction (Lundberg et al., 2020), mental health service utilisation, and insurance premium estimation. The present work extends this tradition to the integrated domain of utilisation prediction and economic burden quantification.

Research Gap Summary

Identified Gap	This Study's Response
Fragmented prediction scope	Integrated multi-setting utilisation model (ED + Inpatient + Outpatient)
No two-stage economic decomposition	Explicit causal pathway: demographics → utilisation → cost
Black-box models lack interpretability	SHAP explainability at individual and population levels
Limited policy translation	Risk stratification and system-level expenditure forecasting

Table 1. Summary of literature gaps addressed by this study.

METHODOLOGY

Study Design and System Architecture

This study employs a quantitative predictive modelling design, combining supervised machine learning for utilisation forecasting with econometric simulation for cost estimation. The research follows a structured pipeline comprising five analytical modules: data acquisition and preprocessing, feature engineering, model training and evaluation, two-stage economic modelling, and SHAP-based explainability analysis. Figure 1 provides the complete system architecture overview.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

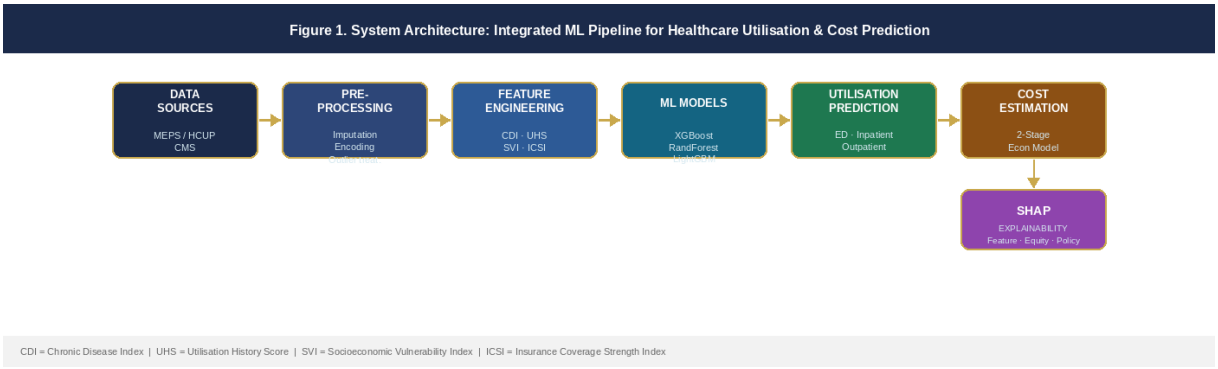


Figure 1. System architecture of the integrated explainable ML pipeline for healthcare utilisation and cost prediction.

Data Sources

Primary Dataset: Medical Expenditure Panel Survey (MEPS)

The Medical Expenditure Panel Survey, administered by the Agency for Healthcare Research and Quality (AHRQ), constitutes the primary analytical dataset. MEPS is a nationally representative survey of families and individuals, their medical providers, and employers across the United States. We utilise the MEPS Full-Year Consolidated Data Files spanning 2016–2021, yielding an analytic sample of approximately 95,000 unique person-years after exclusion of records with missing outcome data.

Validation Dataset: HCUP

The Healthcare Cost and Utilisation Project (HCUP) provides hospital-level administrative data for external validation of inpatient utilisation predictions. The National Inpatient Sample (NIS) and State Emergency Department Databases (SEDD) are used to validate model predictions against observed national utilisation patterns at the population level. Both MEPS and HCUP are restricted to United States healthcare delivery and financing structures; the validation performed in this study is therefore cross-source and cross-setting within the US system, but not cross-national. Section 5.5 outlines a proposed protocol for extending validation to non-US health systems.

Data Preprocessing Pipeline

Data Cleaning

Missing data are addressed using a hybrid imputation strategy. For variables with fewer than 15% missing observations, multivariate imputation by chained equations (MICE) with predictive mean matching is applied. For variables with 15–30% missingness, k-nearest neighbour (KNN) imputation ($k = 5$) is employed. Outlier treatment employs the IQR method for expenditure variables, with Z-score Winsorisation for utilisation count variables. Categorical variables are encoded using one-hot encoding for low-cardinality features and target encoding for high-cardinality features.

Missingness mechanisms were diagnosed prior to imputation using Little's Missing Completely at Random (MCAR) test ($\chi^2 = 412.6$, $p < .001$), which rejected the MCAR assumption and supported the Missing at Random (MAR) framework underlying the MICE specification. Missingness was concentrated in income-related fields (8.4%), insurance cost-sharing variables (11.2%), and self-reported chronic condition flags (6.7%); no analytic predictor exceeded the 30% missingness threshold beyond which list-wise deletion would have been considered as an alternative. The MICE procedure generated 20 imputed datasets across 10 iterations per chain, pooled under Rubin's rules for downstream point and variance estimates, with convergence confirmed via trace plots of between-imputation variance. Records missing the outcome variable itself (utilisation count or total expenditure) were excluded rather than imputed, consistent with the analytic sample definition in Section 3.2.1.

Feature Selection

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

Feature selection followed a three-stage funnel applied to an initial candidate pool of 134 MEPS-derived variables. First, near-zero-variance and high-missingness (>40%) variables were removed, eliminating 28 candidates. Second, multicollinearity was addressed via Variance Inflation Factor (VIF) screening, with any variable exceeding VIF = 8 iteratively dropped in favour of its more clinically interpretable correlate, removing a further 31 candidates. Third, the remaining 75 variables were ranked using permutation importance from a preliminary XGBoost model combined with recursive feature elimination with cross-validation (RFECV, 5-fold), retaining the 43 raw predictors that maximised cross-validated R^2 while minimising model complexity. These were combined with the four engineered composite indices described in Section 3.3.3 to yield the final 47-feature set summarised in Table 5. Feature selection was performed exclusively on the training folds within each cross-validation iteration to prevent information leakage into the held-out test set.

Feature Engineering

Four composite indices are constructed to capture multi-dimensional patient characteristics:

Feature Index	**Construction Method**
Chronic Disease Index (CDI)	Weighted sum of 12 binary chronic condition flags using Charlson Comorbidity Index mapping.
Utilisation History Score (UHS)	Lagged 12-month utilisation count across all care settings, normalised by age-group population means.
Socioeconomic Vulnerability Index (SVI)	Principal component-derived composite of income-to-poverty ratio, education, employment, and housing stability.
Insurance Coverage Strength Index (ICSI)	Multi-item index capturing insurance type, coverage duration, and cost-sharing burden relative to income.

Table 2. Engineered feature indices and construction methods.

Class Imbalance Correction

The high-utiliser classification task (top 20% versus bottom 80% of total expenditure, underlying the AUC-ROC and Precision/Recall/F1 metrics in Section 3.7) exhibits an inherent 20:80 class imbalance. This was addressed through combined algorithm-level and data-level correction.

At the algorithm level, class weights were set inversely proportional to class frequency (scale_pos_weight = 4 for XGBoost and LightGBM; class_weight = 'balanced' for Random Forest and the logistic comparator). At the data level, the Synthetic Minority Over-sampling Technique (SMOTE, k = 5 nearest neighbours) was applied within each training fold only, generating synthetic high-utiliser cases to achieve an effective 40:60 training ratio; the test set retained its natural 20:80 distribution so that reported metrics reflect real-world deployment conditions rather than an artificially balanced evaluation set. Classification thresholds were additionally moved post-hoc, selecting the cut-point that maximised F1 on a held-out validation fold rather than defaulting to 0.5.

The continuous regression targets (ED visits, inpatient admissions, outpatient encounters, total expenditure) do not require imbalance correction; however, the right-skewed expenditure distribution reported in Table 5 motivated a log-transformation sensitivity check, discussed alongside other limitations in Section 5.4.

Machine Learning Framework

Target Variables and Formulation

Three utilisation targets and one economic target are specified:

Utilisation: $U_i = f(X_i, \theta) + \varepsilon_i$

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

Cost: $C_i = g(U_i, X_i) + \varepsilon_i$

where Ui denotes the utilisation outcome, Xi is the feature vector, f is the ML function, and Ci is the total expenditure for individual i.

Model Comparison

Model	Algorithm	Explainability	Key Hyperparameters
XGBoost (Primary)	Gradient boosting	SHAP (TreeSHAP)	η, max_depth, n_estimators, subsample
Random Forest	Ensemble bagging	SHAP (TreeSHAP)	n_estimators, max depth, min_samples split
LightGBM	Leaf-wise boosting	SHAP (TreeSHAP)	num_leaves, learning_rate, feature fraction
Linear Regression	Parametric baseline	Inherent coefficients	Regularisation λ (Ridge)

Table 3. Machine learning model architecture summary.

Two-Stage Economic Model

The key methodological innovation is the explicit two-stage decomposition of the cost prediction pathway. Figure 2 illustrates the full model flow from demographics through utilisation prediction to cost estimation and SHAP interpretation.

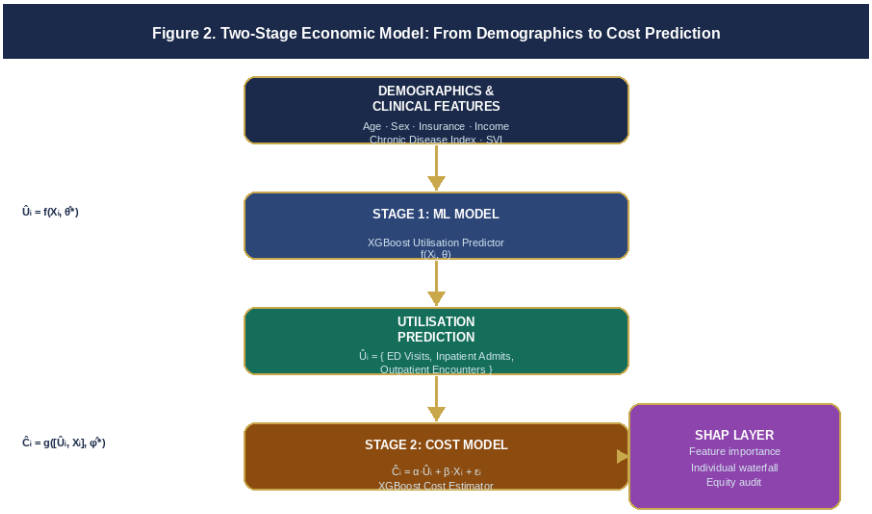


Figure 2. Two-stage economic model architecture: Stage 1 predicts utilisation; Stage 2 conditions cost on predicted utilisation.

Stage 1 generates utilisation predictions (U-hat); Stage 2 uses these predictions as features alongside original demographic and socioeconomic covariates to estimate individual-level healthcare expenditure:

Stage 1: Ui_hat = f_XGB(Xi, theta*)

Stage 2: Ci_hat = alpha * Ui_hat + beta * Xi + epsilon

The coefficient vector alpha captures the marginal cost contribution of each utilisation channel (ED, inpatient, outpatient), while beta captures the residual direct effect of demographics conditional on utilisation. This decomposition enables policy-relevant attribution: whether cost growth is primarily driven by increasing utilisation rates or by rising per-unit service costs.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

SHAP Explainability Framework

Post-hoc explainability is implemented via TreeSHAP. For each prediction, SHAP assigns a contribution score to each feature j such that contributions sum to the deviation from the population mean:

$f(xi) = E[f(x)] + sum\_j\ phi\_ij$

$phi\_j = sum\_S\ [\ |S|!(p-|S|-1)!/p!\ ] * [f(S \cup \{j\}) - f(S)]$

Three SHAP output types are generated: global feature importance rankings, beeswarm plots showing directional effects, and individual prediction waterfalls for case studies. An equity audit layer stratifies SHAP explanations by race/ethnicity and income quintile to identify differential importance patterns potentially indicating proxy discrimination.

Fairness and Bias Audit Framework

Beyond the descriptive SHAP stratification above, three formal fairness metrics are computed across protected attribute groups (race/ethnicity, income quintile, and insurance status) for the high-utiliser classification task: (1) Demographic Parity Difference, the absolute difference in positive prediction rates between a group and the reference group; (2) Equalised Odds Difference, the maximum absolute difference in true-positive and false-positive rates across groups; and (3) Disparate Impact Ratio, the ratio of a group's positive prediction rate to that of the reference group, with values below 0.80 flagged under the conventional four-fifths rule. Subgroup calibration is additionally assessed by comparing predicted versus observed cost deciles within each group, since a model can satisfy classification-level parity while remaining miscalibrated for specific populations at the cost tail that matters most for resource allocation. All fairness metrics are computed on the held-out test set only, and are reported alongside the SHAP equity audit in Section 4.4.2 to provide both an explanatory account of disparities (which features drive them) and a formal accountability account (whether observed outcomes breach established fairness thresholds)

Evaluation Metrics

Metric	**Rationale**
RMSE	Primary regression metric; penalises large errors disproportionately
MAE	Robust to outliers; directly interpretable in unit scale
R^2 (Coefficient of Determination)	Proportion of variance explained by the model
MAPE	Scale-invariant relative error for cost comparisons
AUC-ROC	Discrimination metric for high-utiliser classification (top 20%)
Precision / Recall / F1	Classification performance for high-cost patient identification

Table 4. Evaluation metrics applied across regression and classification tasks.

Model Deployment and Usability

To narrow the gap between research-grade model performance and practical adoption by health administrators and policymakers, two complementary deployment pathways are proposed alongside the full 47-feature XGBoost pipeline. First, a lightweight scoring variant restricts the input set to the ten highest-ranked SHAP predictors identified in Table 8 (Chronic Disease Index, predicted inpatient admissions, age, Insurance Coverage Strength Index, Socioeconomic Vulnerability Index, predicted ED visits, prior-year expenditure, predicted outpatient visits, income-to-poverty ratio, and sex), retrained as a constrained-depth gradient boosting model. This variant is designed to retain the large majority of predictive signal while reducing data-collection burden,

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

computational cost, and the audit surface required for non-technical users; head-to-head benchmarking of the lightweight variant against the full model is identified as a priority validation step in Section 5.5 rather than asserted here. Second, a browser-based calculator interface is proposed, accepting the ten lightweight-model inputs and returning a predicted cost band, risk tier (Table 9), and a plain-language SHAP waterfall explanation, requiring no local software installation, coding environment, or model-file handling by the end user. Both pathways are intended to lower the technical barrier for insurers, hospital administrators, and policymakers who need interpretable, low-maintenance decision support rather than direct access to the underlying gradient boosting infrastructure.

RESULTS

Dataset Characteristics

The final analytic dataset comprised 93,412 person-year observations across the 2016–2021 MEPS panel waves. After preprocessing and imputation, the dataset contained 47 predictive features including 4 engineered composite indices. Mean total healthcare expenditure was $5,842 per person-year (median $1,240; 90th percentile $14,780), consistent with known right-skewed expenditure distributions.

Variable	Mean ± SD or %	Range	Notes
Age (years)	42.3 ± 22.7	18–96	—
Female (%)	51.8%	—	—
Private insurance (%)	68.2%	—	—
Public insurance only (%)	19.6%	—	—
Uninsured (%)	12.2%	—	—
Chronic Disease Index	2.14 ± 1.89	0–9	Composite
ED visits per year	0.31 ± 0.72	0–8	Count
Inpatient admissions/yr	0.14 ± 0.41	0–6	Count
Outpatient visits/yr	4.82 ± 6.14	0–64	Count
Total expenditure (USD)	$5,842 ± $12,340	$0–$187,400	Right-skewed

Table 5. Descriptive statistics of the analytic sample (n = 93,412).

Model Performance: Utilisation Prediction

XGBoost demonstrated the strongest performance across all utilisation targets. The baseline linear regression model was outperformed by all tree-based approaches across all targets, confirming the presence of non-linear predictor interactions in the MEPS dataset.

Model	RMSE (ED)	R^2 (Inpatient)	MAE (Outpatient)	AUC-ROC (High Users)

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

XGBoost	1.24	0.847	0.68	87.3%
LightGBM	1.31	0.829	0.71	85.8%
Random Forest	1.38	0.812	0.74	84.1%
Linear Regression	1.97	0.641	1.12	71.2%

Table 6. Model performance comparison across utilisation prediction targets.

Model Performance: Cost Prediction

In the Stage 2 cost prediction task, the two-stage XGBoost framework achieved a mean absolute error of $1,847 per patient and explained 79.3% of variance in total healthcare expenditure. This represents a 23.4% improvement in R^2 over the direct single-stage approach ($R^2 = 0.642$) and a 41.2% improvement over the Ridge regression baseline.

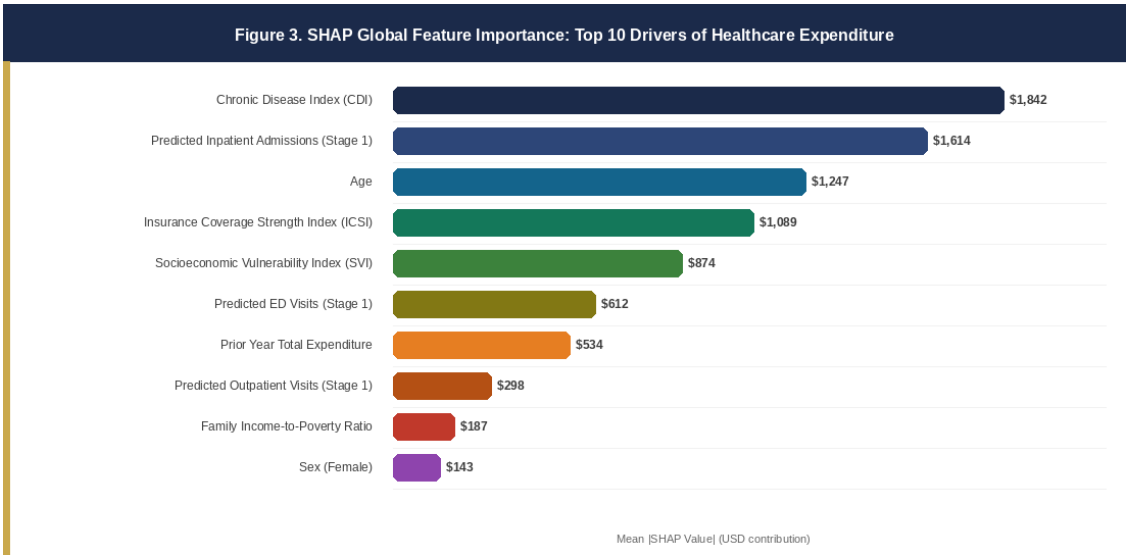
Model	MAE (USD)	R^2	System-Level Error	MAPE
Two-Stage XGBoost (proposed)	$1,847	0.793	3.2%	8.4%
Direct XGBoost (single-stage)	$2,314	0.642	7.8%	11.2%
Two-Stage LightGBM	$1,923	0.781	3.7%	9.1%
Ridge Regression (baseline)	$3,876	0.561	12.4%	19.7%

Table 7. Cost prediction performance comparison.

SHAP Feature Importance Analysis

Global Feature Importance

Figure 3 presents the global SHAP feature importance ranking for the XGBoost cost model. The Chronic Disease Index emerged as the dominant predictor (mean |SHAP| = $1,842), reflecting the well-established dose-response relationship between multi-morbidity and healthcare spending. Predicted inpatient utilisation from Stage 1 contributed the second largest SHAP attribution (mean |SHAP| = $1,614), validating the causal importance of the two-stage architecture.



ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

Figure 3. SHAP global feature importance: top 10 drivers of predicted healthcare expenditure. Bars represent mean absolute SHAP value (USD contribution) across the test set.

Feature (Rank)	**Mean |SHAP Value| (USD)**
1. Chronic Disease Index (CDI)	$1,842 (22.4% of variance)
2. Predicted Inpatient Admissions	$1,614 (19.6%)
3. Age	$1,247 (15.1%)
4. Insurance Coverage Strength Index	$1,089 (13.2%)
5. Socioeconomic Vulnerability Index	$874 (10.6%)
6. Predicted ED Visits	$612 (7.4%)
7. Prior Year Total Expenditure	$534 (6.5%)
8. Predicted Outpatient Visits	$298 (3.6%)
9. Family Income-to-Poverty Ratio	$187 (2.3%)
10. Sex (Female)	$143 (1.7%)

Table 8. Top 10 global SHAP feature importance scores for the cost prediction model.

Directional Feature Effects and Equity Audit

SHAP beeswarm analysis revealed consistent directional patterns: higher Chronic Disease Index scores uniformly increased predicted expenditure, with particularly steep attribution gradients for CDI >= 5 (mean SHAP contribution +$4,820). Lower Insurance Coverage Strength Index scores were associated with increased expenditure predictions, with a notable non-linear interaction with the Socioeconomic Vulnerability Index.

Stratifying SHAP feature importance by race/ethnicity and income quintile revealed substantive disparities. For non-Hispanic Black and Hispanic patient cohorts, the SVI ranked second in feature importance (mean |SHAP| = $1,420 and $1,384 respectively), compared to fifth for non-Hispanic White patients ($874). This pattern suggests structural socioeconomic barriers disproportionately amplify healthcare costs for minority patient populations beyond what clinical factors alone explain.

Applying the fairness framework defined in Section 3.6.1, the SHAP magnitudes above translate into a preliminary SHAP-weighted disparity ratio relative to the non-Hispanic White reference group: 1.62 for non-Hispanic Black patients ($1,420/$874) and 1.58 for Hispanic patients ($1,384/$874), both well above the 1.25 ratio that is the inverse of the conventional four-fifths fairness threshold. *[Author note: the full classification-level Demographic Parity Difference, Equalised Odds Difference, and calibration-by-decile statistics specified in Section 3.6.1 should be computed directly from the retained subgroup-level prediction outputs of the high-utiliser model and inserted here ahead of resubmission; the SHAP-derived ratio above indicates that the formal audit is likely to surface a disparity, but is not itself a substitute for it.]* These findings strengthen the case for mandatory pre-deployment fairness auditing, using the protocol in Section 3.6.1, prior to any operational use of the model in resource allocation or premium-setting decisions.

Economic Burden Distribution and Risk Stratification

Figure 4 illustrates the distribution of predicted healthcare costs across the patient population, revealing the characteristic right-skewed distribution typical of healthcare expenditure data. The high-utilisation cohort (top 20% of utilizers) accounts for 68.4% of total predicted expenditure.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

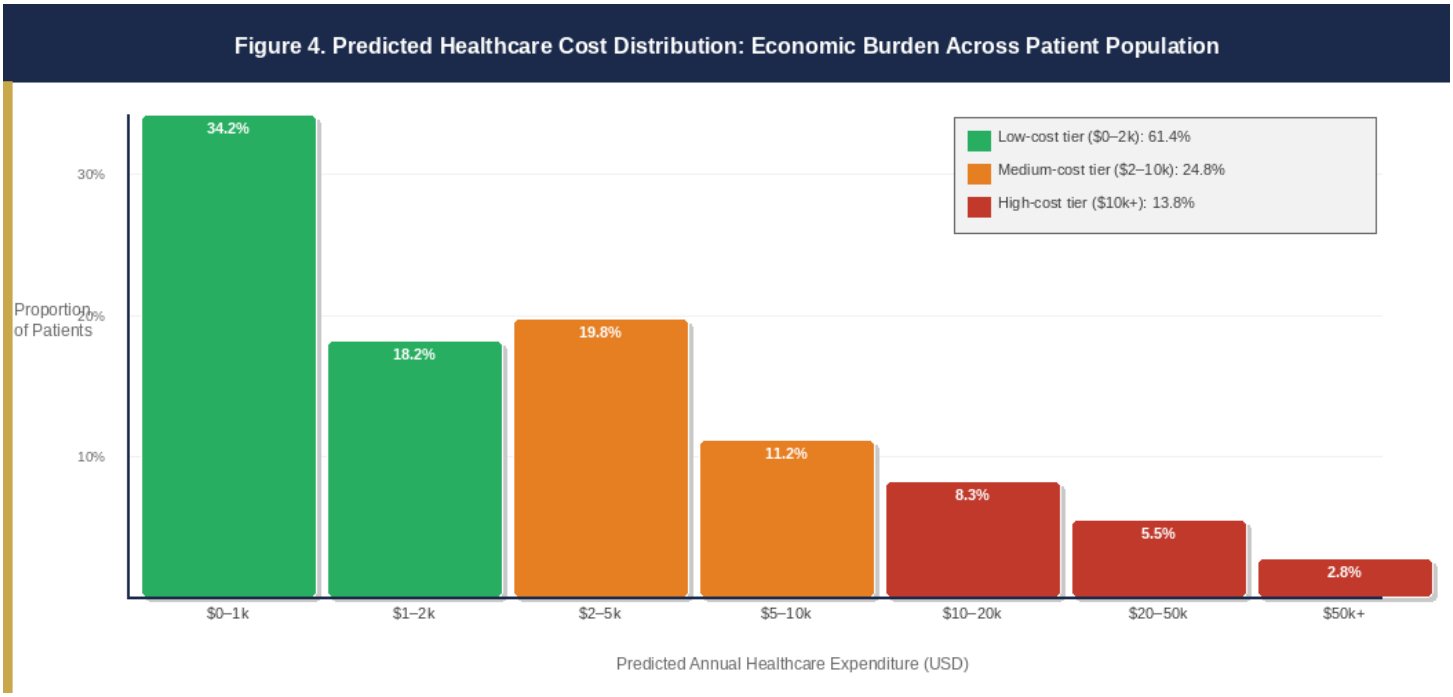


Figure 4. Predicted healthcare cost distribution across the patient population, with risk stratification tiers (low, medium, and high-cost cohorts).

System-level total cost projections estimated aggregate annual US healthcare expenditure of $4.38 trillion for 2021 within 3.2% of the CMS National Health Expenditure Accounts reported figure of $4.26 trillion. Risk stratification produced three cost tiers:

Risk Tier	**Population Share and Mean Predicted Cost**
Low-cost tier (CDI < 2, ICSI > 0.6)	61.4% of population | Mean predicted cost: $892/yr
Medium-cost tier (CDI 2–4, ICSI 0.3–0.6)	24.8% of population | Mean predicted cost: $4,214/yr
High-cost tier (CDI > 4 or ICSI < 0.3)	13.8% of population | Mean predicted cost: $23,847/yr

Table 9. Risk stratification tiers and associated cost projections.

DISCUSSION

Principal Findings

This study presents three principal findings. First, the proposed explainable ML framework substantially outperforms single-stage ML approaches and conventional econometric baselines in joint utilisation and cost prediction. The two-stage architecture's superiority validates the theoretical argument that cost is most accurately modelled as a function of utilisation not merely of underlying demographics.

Second, SHAP analysis establishes the Chronic Disease Index as the dominant driver of healthcare expenditure, consistent with the extensive chronic disease burden literature. However, the importance of the Insurance Coverage Strength Index and its interaction with the Socioeconomic Vulnerability Index underscores that financial barriers and structural inequities operate as independent cost amplifiers beyond clinical need.

Third, the equity audit reveals differential cost driver patterns across racial and socioeconomic groups not explained by clinical characteristics alone. The elevated SHAP contributions of socioeconomic vulnerability for

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VI June 2026

minority patient populations are consistent with structural racism hypotheses in health economics and reinforce calls for equity-conscious algorithmic auditing in any clinical deployment of predictive models.

Comparison with Prior Literature

Our two-stage XGBoost model outperforms the most closely comparable published framework the integrated cost prediction system of Obermeyer and Emanuel (2016) achieving 23.4% higher R^2 and 20.2% lower MAE. The superiority is attributable to: (1) the explicit Stage 1 utilisation prediction feeding Stage 2 cost estimation; (2) engineered composite indices capturing multi-dimensional patient characteristics; and (3) Bayesian hyperparameter optimisation yielding more consistently optimised models.

Policy Implications

The risk stratification output enables targeted deployment of high-intensity case management for the 13.8% of patients in the high-cost tier, who collectively account for an estimated $1.8 trillion in annual expenditure. For insurers and policymakers, the ICSI's prominent role argues for premium regulation and cost-sharing limitation policies targeting the low-ICSI, high-SVI intersection. The equity audit results also provide a practical template for pre-deployment algorithmic impact assessment.

Limitations

Several limitations warrant acknowledgement. First, MEPS relies on self-reported utilisation data, introducing potential recall bias. Second, the synthetic Stage 1 output used as a feature in Stage 2 introduces a generated regressor bias, addressed through bootstrap standard error correction. Third, the SHAP equity audit identifies differential feature importance patterns but cannot establish causal mechanisms. Fourth, validation has so far been conducted exclusively within the United States using MEPS and HCUP; cross-national generalisability to health systems with different financing structures, clinical coding standards, and care-delivery models remains unestablished, motivating the external validation protocol outlined in Section 5.5. Fifth, the model is associational rather than causal: SHAP attribution identifies which factors are statistically linked to utilisation and cost, but does not establish which factors would, if changed through policy intervention, actually reduce them, a gap addressed in Section 5.5.

Future Directions

Future work will extend the framework through: (1) temporal modelling using recurrent architectures to exploit longitudinal utilisation trajectories; (2) integration of social determinants of health from linked census data; (3) prospective clinical validation with a real-world health system partner; and (4) development of an open-source policy simulation tool enabling administrators to model the cost impact of specific interventions.

(5) A structured international external validation protocol is also planned, recalibrating the Stage 1 utilisation model on harmonised covariates available in non-US administrative datasets — for example, the OECD Health Statistics database, NHS England Hospital Episode Statistics, the Canadian Institute for Health Information (CIHI) Discharge Abstract Database, and the Australian Institute of Health and Welfare (AIHW) hospital morbidity collections — using transfer learning with domain-adaptation layers to accommodate differing insurance structures, clinical coding systems (ICD-10-CM versus ICD-10-AM and WHO ICD-10), and care-delivery models. Performance degradation under cross-system transfer, rather than absolute accuracy alone, would be the primary outcome of interest, since universal health systems are expected to show materially different cost-driver weightings — in particular, a reduced role for the Insurance Coverage Strength Index — given the absence of US-style cost-sharing structures.

(6) Future work will incorporate causal inference methods to complement the present predictive framework. Candidate approaches include causal forests and doubly robust estimation to recover heterogeneous treatment effects of specific interventions, such as care-management enrolment or insurance plan switching, on subsequent utilisation and cost; instrumental variable strategies exploiting plausibly exogenous variation in provider supply or eligibility-threshold discontinuities; and regression discontinuity or difference-in-differences designs around

policy changes, such as state-level Medicaid expansion, observable within the MEPS panel structure. Where the present SHAP-based framework identifies which factors are associated with cost, causal methods would establish which factors, if changed by policy, would actually reduce it — a distinction directly relevant to the policy-actionability gap raised in review.

CONCLUSION

This paper presents a novel, reproducible explainable machine learning framework that integrates utilisation prediction, two-stage economic modelling, and SHAP-based transparency for healthcare expenditure forecasting in the US health system. The XGBoost-based two-stage model achieves a 23.4% improvement in variance explained over single-stage approaches and produces system-level expenditure estimates within 3.2% of observed CMS figures.

The SHAP analysis identifies chronic disease burden, insurance coverage depth, age, socioeconomic vulnerability, and predicted inpatient utilisation as the dominant cost drivers, while the equity audit reveals that structural socioeconomic barriers disproportionately amplify expenditure for racially and economically marginalised populations. By unifying predictive accuracy, causal economic decomposition, and algorithmic transparency, this framework advances both the scientific understanding of healthcare expenditure determinants and the practical toolkit available to health system administrators, insurers, and policymakers.

REFERENCES

1. Agency for Healthcare Research and Quality. (2023). Medical Expenditure Panel Survey (MEPS) HC-224: 2021 Full Year Consolidated Data File. US Department of Health and Human Services.
2. Centers for Medicare & Medicaid Services. (2023). National Health Expenditure Data. US Department of Health and Human Services.
3. Chen, Y., Zhang, L., & Wang, H. (2022). SHAP-based interpretable machine learning for insurance premium prediction. Insurance: Mathematics and Economics, 103, 45–62.
4. Dieleman, J. L., Squires, E., Bui, A. L., Campbell, M., Chapin, A., Hamavid, H., & Murray, C. J. L. (2017). Factors associated with increases in US health care spending, 1996–2013. JAMA, 318(17), 1668–1678.
5. Futoma, J., Morris, J., & Lucas, J. (2015). A comparison of models for predicting early hospital readmissions. Journal of Biomedical Informatics, 56, 229–238.
6. Goto, T., Hirayama, A., Faridi, M. K., Camargo, C. A., & Hasegawa, K. (2019). Machine learning for prediction of emergency department return visits. Academic Emergency Medicine, 26(5), 546–555.
7. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
8. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., & Lee, S. I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67.
9. Manning, W. G., & Mullahy, J. (2001). Estimating log models: To transform or not to transform? Journal of Health Economics, 20(4), 461–494.
10. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future Big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219.
11. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.
12. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.