

Predictive Analytics for Healthcare Cost Reduction

Predictive Analytics for Healthcare Cost Reduction: Using Artificial Intelligence to Identify Preventable High-Cost Patients Before Crisis Care

MIS-690 Capstone Project Thesis

Submitted to Grand Canyon University

Graduate Faculty of the Colangelo College of Business

in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Business Analytics

by

Sara Seri

Phoenix, Arizona

July 2026

Abstract

This capstone project developed and evaluated machine learning models to identify individuals likely to become high-cost healthcare users before crisis-level utilization occurs. The study used the Agency for Healthcare Research and Quality Medical Expenditure Panel Survey (MEPS) HC-252 Panel 27 longitudinal file. Information available in 2022 was used to predict whether a person entered the weighted top decile of total healthcare expenditures in 2023. The analytic sample contained 8,175 individuals, and the weighted high-cost threshold was \$17,749. Logistic Regression, Decision Tree, Random Forest, and XGBoost models were evaluated across 70/30, 80/20, and 95/5 train-test partitions, followed by evaluation on a sealed 10% validation set. Random Forest produced the strongest average precision-recall performance and was selected as the final model. On sealed validation data, it achieved ROC-AUC 0.817, PR-AUC 0.443, recall 0.604, precision 0.472, and top-decile lift 4.31. Prior-year expenditure, prescription volume, general health, walking limitation, and age were among the strongest predictors. The findings demonstrate that longitudinal predictive analytics can meaningfully stratify future cost risk and may support targeted care management, but the model should be deployed as decision support with clinical review, fairness monitoring, calibration checks, and continuous validation.

Keywords: predictive analytics, healthcare expenditures, high-cost patients, machine learning, MEPS, random forest, population health

Table of Contents

- Abstract
- Business Problem Identification
- Background
- Business Problem Statement
- Analytics Assumptions
- Analytics Problem Statement
- Data Understanding, Acquisition, and Preprocessing
- Collection of Initial Data
- Description of Data
- Exploration of Data
- Verify Data Quality
- Hypothesis Statements
- Data Diagnostics and Descriptive Summary
- Methodology Approach and Model Building
- Model Evaluation
- External Model Verification and Calibration
- Model Deployment and Model Life Cycle
- Recommendations
- Conclusions
- References
- Appendix A: Variable Dictionary

Predictive Analytics for Healthcare Cost Reduction

Healthcare organizations face sustained pressure to improve outcomes while controlling expenditures. A relatively small group of patients accounts for a disproportionate share of spending, yet high-cost status is not always persistent from one year to the next. This creates a difficult operational problem: care-management resources are limited, but intervening only after hospitalization or emergency care occurs is often too late to prevent the most expensive outcomes. This project addresses that challenge by using longitudinal data and machine learning to identify individuals at elevated risk of becoming high-cost patients during the following year.

Business Problem Identification

Background

National healthcare spending has continued to grow, while person-level expenditures remain highly concentrated. AHRQ reported that in 2022 the top 5% of people accounted for approximately half of healthcare expenses, demonstrating why high-cost patient identification is strategically important. Health systems, payers, accountable care organizations, and public programs therefore need reliable approaches for identifying patients who may benefit from preventive outreach, medication management, care coordination, transportation assistance, behavioral health support, or other interventions before costly deterioration occurs.

Traditional rule-based methods often depend heavily on prior spending or a limited number of diagnoses. These approaches are interpretable but may overlook interactions between functional limitations, self-reported health, utilization, chronic conditions, socioeconomic factors, and insurance status. Predictive analytics provides a structured way to integrate these factors and assign prospective risk scores that support population-level prioritization.

Business Problem Statement

Healthcare organizations experience preventable financial and operational pressure when individuals at risk of becoming high-cost patients are identified only after emergency department use, hospitalization, or advanced disease progression. Existing reactive approaches limit opportunities for earlier intervention and may lead to inefficient allocation of care-management resources. A prospective analytics solution is needed to identify elevated future cost risk using information available before the high-cost period begins.

Analytics Assumptions

- Historical utilization, health status, and expenditure patterns contain information relevant to future cost risk.
- The HC-252 longitudinal survey data are sufficiently representative for population-level analysis when survey weights are applied.
- The model is intended for prioritization and decision support, not autonomous clinical decision-making.
- A high-cost label based on the weighted top decile is operationally useful but does not imply that every high-cost event is preventable.
- False negatives are particularly important because they represent high-cost patients who may not be offered preventive support.

Analytics Problem Statement

The objective is to develop supervised classification models that use 2022 demographic, socioeconomic, health-status, chronic-condition, prior-utilization, and expenditure information to predict whether an individual will be in the weighted top 10% of total healthcare expenditures in 2023. The models will be evaluated for discrimination, minority-class performance, calibration, ranking effectiveness, stability across train-test partitions, and operational usefulness.

Data Understanding, Acquisition, and Preprocessing

Collection of Initial Data

The study used the AHRQ MEPS HC-252 Panel 27 Longitudinal Public Use File, which links information for respondents participating during 2022 and 2023. The original Excel file contained 8,292 records and 2,648 variables. After requiring a valid 2023 expenditure outcome and a positive longitudinal survey weight, the analytic sample contained 8,175 individuals. The longitudinal structure allows predictors from 2022 to be separated from the 2023 outcome, reducing temporal leakage and supporting a prospective research design.

Description of Data

Table 1. Selected Data Classification

Variable	Description	Type	Role
TOTEXPY2	2023 total healthcare expenditure	Numeric	Outcome source
HIGH_COST_Y2	Weighted top-decile 2023 expenditure indicator	Binary	Target
TOTEXPY1	2022 total healthcare expenditure	Numeric	Predictor
AGEY1X	Age in 2022	Numeric	Predictor
SEX	Sex	Categorical	Predictor
RACETHX	Race/ethnicity category	Categorical	Predictor
POVCATY1	2022 poverty category	Ordinal	Predictor
INSCOVY1	2022 insurance coverage	Categorical	Predictor
RTHLTH3	Self-reported general health	Ordinal	Predictor
MNHLTH3	Self-reported mental health	Ordinal	Predictor
ERTOTY1	2022 emergency visits	Count	Predictor
IPDISY1	2022 inpatient discharges	Count	Predictor
RXTOTY1	2022 prescription events	Count	Predictor
LONGWT	Longitudinal survey weight	Numeric	Estimation weight

Target Construction and Leakage Prevention

The high-cost target was defined using the weighted 90th percentile of 2023 total expenditure. The calculated threshold was \$17,749.33. Individuals with TOTEXPY2 at or above this value were assigned HIGH_COST_Y2 = 1. All 2023 utilization, health-status, and expenditure variables were excluded from model predictors. Only baseline demographics and information observed in 2022 were permitted. This separation is central to the validity of the “before crisis care” objective.

Feature Engineering

- Log-transformed prior expenditure to reduce extreme right skew.
- Chronic disease burden score calculated from 11 diagnosed-condition indicators.
- Utilization index combining prior emergency, inpatient, office, outpatient, prescription, and home-health use.
- Binary functional-limitation indicators for walking, activities of daily living, and instrumental activities of daily living.
- Log-transformed family income to reduce skewness.

Verify Data Quality

MEPS uses negative values to represent inapplicability, nonresponse, or other special codes. These values were not interpreted as substantive measurements. Continuous predictors were set to missing where appropriate and imputed with training-set medians within each model pipeline. Categorical predictors were imputed using the most frequent training category. Condition variables were converted to affirmative diagnosis indicators. No outcome observations with invalid 2023 total expenditure or nonpositive longitudinal weights were retained.

Table 2. Highest Missingness Among Analytic Variables

Variable	Missing_Count	Missing_Percent
employed	1512	18.4954
education_years	534	6.5321

Predictive Analytics for Healthcare Cost Reduction

Variable	Missing_Count	Missing_Percent
home_health_days	94	1.1498
iadl_limitation	87	1.0642
adl_limitation	83	1.0153
walking_limitation	79	0.9664
mental_health	79	0.9664
general_health	75	0.9174
log_family_income	74	0.9052
family_income	74	0.9052
TOTEXPY1	66	0.8073
outpatient_visits	66	0.8073
prescription_count	66	0.8073
inpatient_discharges	66	0.8073
er_visits	66	0.8073

Hypothesis Statements

H01: Demographic, socioeconomic, health-status, chronic-condition, prior-utilization, and prior-expenditure factors do not predict high-cost status in the following year.

H11: At least one of these factors provides meaningful predictive information about future high-cost status.

H02: Nonlinear machine learning models do not outperform Logistic Regression on ranking and discrimination metrics.

H12: At least one nonlinear model outperforms Logistic Regression on prospective high-cost prediction.

H03: Prior healthcare utilization and expenditure do not improve identification of future high-cost patients.

H13: Prior healthcare utilization and expenditure improve identification of future high-cost patients.

Data Diagnostics and Descriptive Summary

Table 3. Dataset Summary

Statistic	Value
Unweighted observations	8,175.00
Weighted 2023 expenditure p90 threshold	17,749.33
High-cost observations	963.00
High-cost share (unweighted)	0.1178
Mean 2022 expenditure	7,733.01
Median 2022 expenditure	1,797.00
Mean 2023 expenditure	8,456.16
Median 2023 expenditure	1,747.00

Table 4. Weighted Summary

Metric	Value
Weighted high-cost share	0.1000
Weighted mean 2022 expenditure	6,604.82
Weighted mean 2023 expenditure	7,328.55
Weighted p90 threshold	17,749.33

Table 5. Descriptive Statistics

Variable	count	mean	std	min	25%	50%	75%	max
age	8,109.00	43.3150	23.7498	0.0000	23.0000	44.0000	63.0000	85.0000
education_years	7,641.00	12.0716	4.4064	0.0000	11.0000	12.0000	16.0000	17.0000

Predictive Analytics for Healthcare Cost Reduction

Variable	count	mean	std	min	25%	50%	75%	max
family_income	8,101.00	91,449.08	84,543.27	0.0000	32,515.00	68,888.00	125,462.00	674,227.00
general_health	8,100.00	2.3006	1.0365	1.0000	1.0000	2.0000	3.0000	5.0000
mental_health	8,096.00	2.2032	1.0281	1.0000	1.0000	2.0000	3.0000	5.0000
chronic_burden	8,175.00	1.3470	1.6862	0.0000	0.0000	1.0000	2.0000	10.0000
er_visits	8,109.00	0.2098	0.6678	0.0000	0.0000	0.0000	0.0000	16.0000
inpatient_discharges	8,109.00	0.0805	0.3523	0.0000	0.0000	0.0000	0.0000	5.0000
office_visits	8,109.00	7.0437	13.3033	0.0000	1.0000	3.0000	8.0000	282.00
outpatient_visits	8,109.00	1.0582	4.6913	0.0000	0.0000	0.0000	1.0000	160.00
prescription_count	8,109.00	10.1151	18.4571	0.0000	0.0000	2.0000	13.0000	241.00
home_health_days	8,081.00	2.9798	25.0565	0.0000	0.0000	0.0000	0.0000	402.00
TOTEXPY1	8,109.00	7,733.01	23,253.54	0.0000	330.00	1,797.00	6,635.00	802,849.00
TOTEXPY2	8,175.00	8,456.16	22,574.44	0.0000	259.00	1,747.00	7,254.00	574,675.00

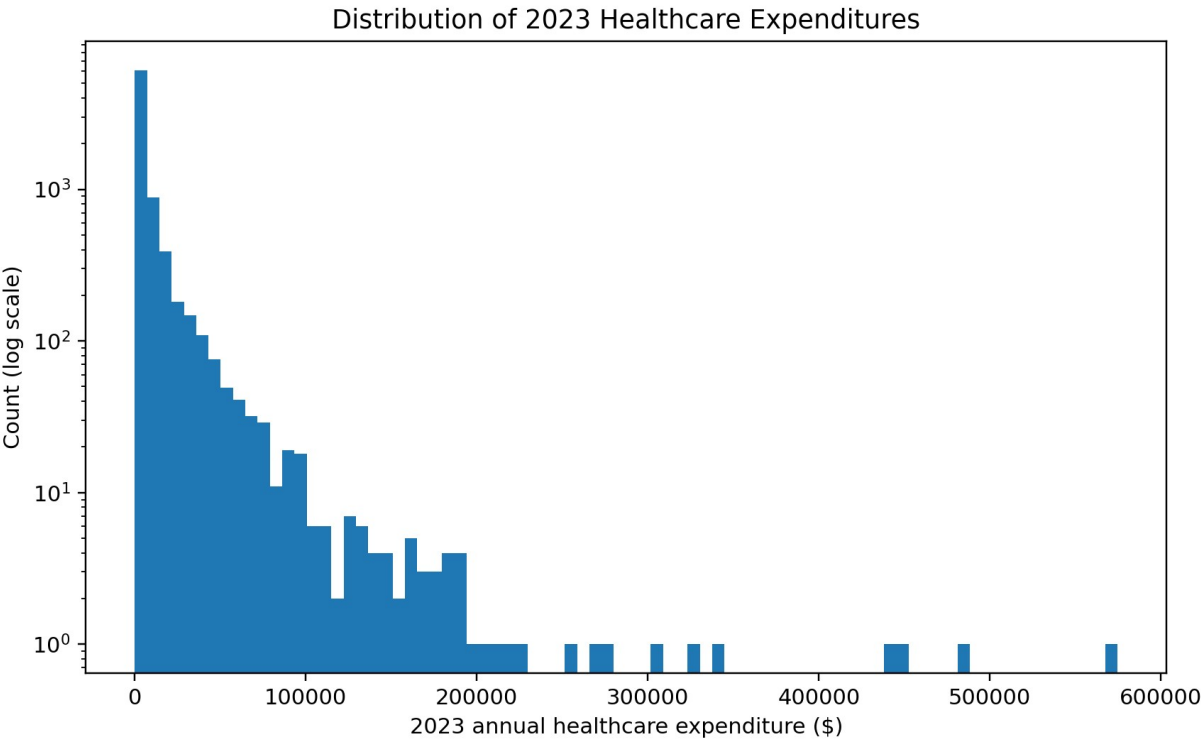


Figure 1. Distribution of 2023 healthcare expenditures.

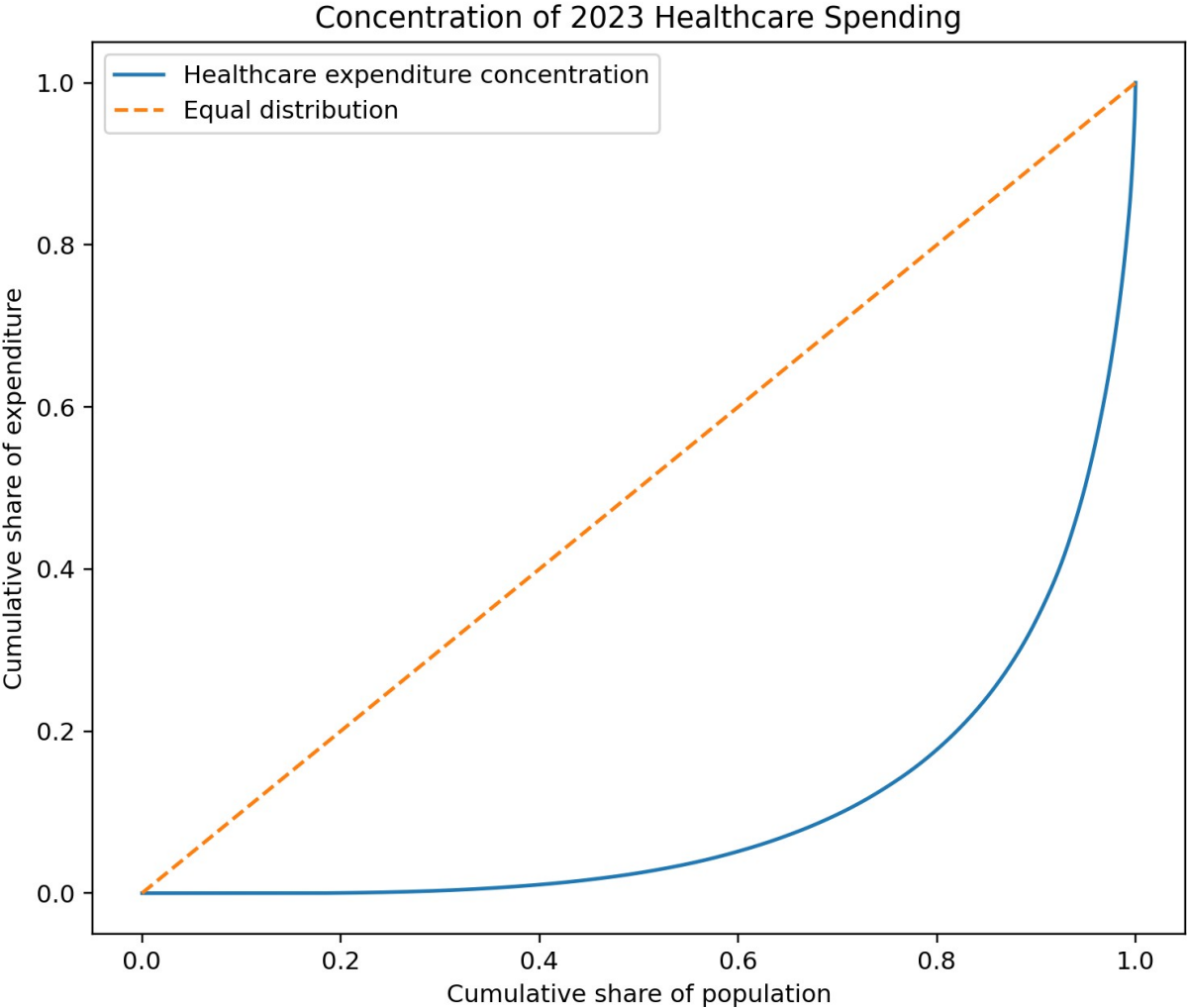


Figure 2. Weighted concentration curve for 2023 healthcare spending.

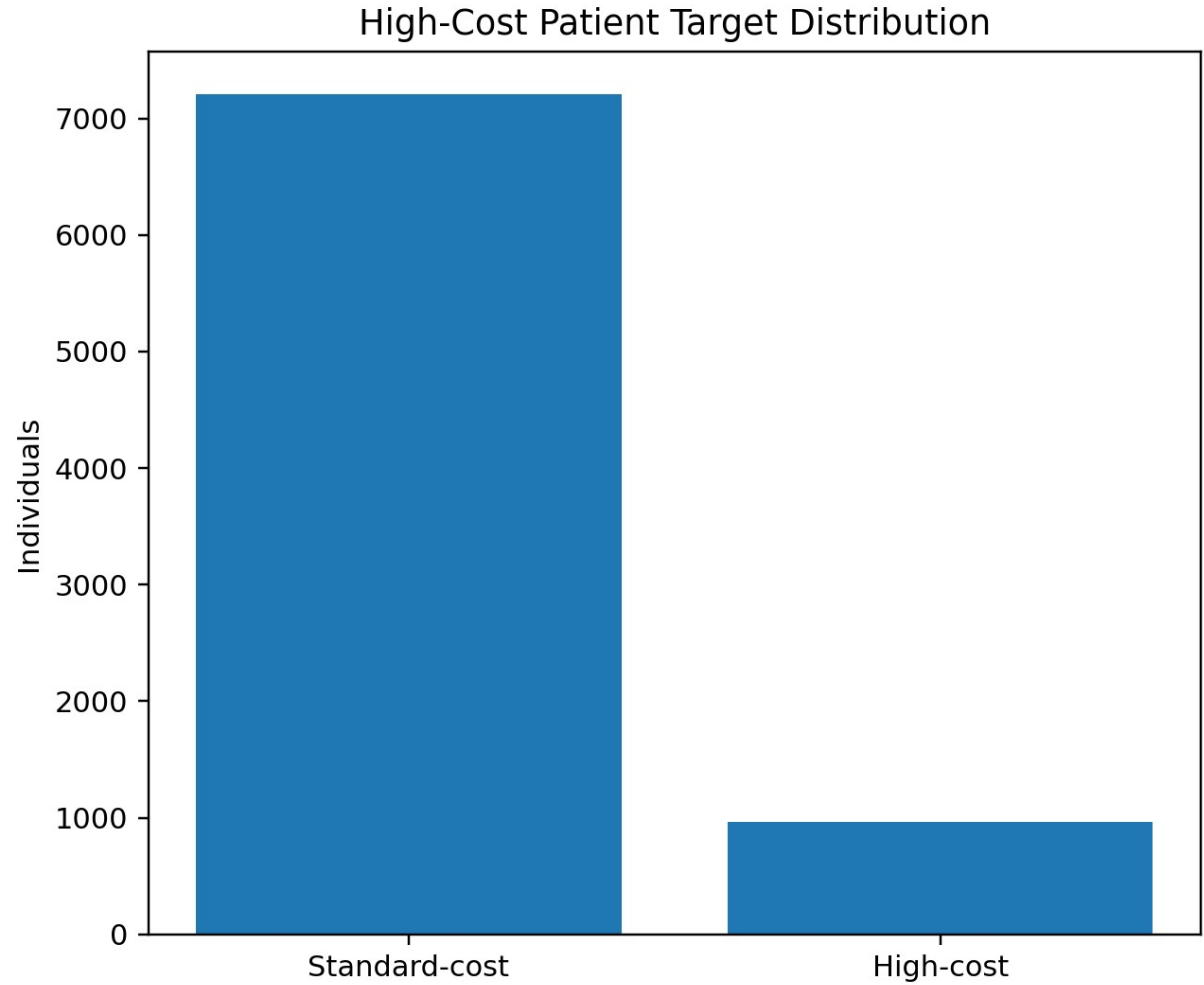


Figure 3. Unweighted class distribution for the high-cost target.



Figure 4. Relationship between prior-year and future healthcare spending.

High-cost individuals had mean 2023 expenditures of \$48,895, compared with \$3,056 among standard-cost individuals. This gap confirms that the target separates a financially distinct group. However, prior spending and future spending are not perfectly aligned, which supports the need for a multivariable prospective model rather than a simple rule based exclusively on past cost.

Methodology Approach and Model Building

Modeling Methods

Four supervised classification algorithms were developed. Logistic Regression provided an interpretable baseline and used balanced class weights. Decision Tree supplied transparent nonlinear rules. Random Forest averaged multiple trees to model interactions and reduce single-tree variance. XGBoost sequentially combined shallow trees and used class weighting to address the minority outcome. Preprocessing was performed inside each model pipeline so that imputation, standardization, and categorical encoding were learned only from training data.

Test Design

Ten percent of the analytic sample was sealed before model comparison and used only for final validation. The remaining development sample was evaluated using 70/30, 80/20, and 95/5 stratified train-test partitions. Identical random seeds and outcome stratification were used to support comparability. The 95/5 partition was retained because it appears in the professor-provided benchmark, but its small test sample was interpreted cautiously due to higher metric variability.

Model Tuning

Table 6. Model Settings

Model	Final settings
Logistic Regression	C = 0.5; balanced class weights; liblinear solver
Decision Tree	Maximum depth 5; minimum leaf size 25; balanced weights
Random Forest	220 trees; maximum depth 12; minimum leaf size 5; square-root feature sampling
XGBoost	250 trees; learning rate .04; depth 3; minimum child weight 5; subsampling .85; L2 regularization 5

Model Assessment

Table 7. Model Performance Across Train-Test Partitions

Split	Model	Accuracy	Precision	Recall	Specificity	F1	Balance_d_Accuracy	ROC_AUC	PR_AUC	Top_De cile_Capture	Top_De cile_Lift
70/30	Logistic Regression	0.7649	0.3035	0.7692	0.7644	0.4353	0.7668	0.8536	0.4628	0.4577	4.5936
70/30	Decision Tree	0.7971	0.3374	0.7500	0.8034	0.4654	0.7767	0.8422	0.4378	0.4269	4.2848
70/30	Random Forest	0.8668	0.4530	0.6308	0.8984	0.5273	0.7646	0.8609	0.4944	0.4731	4.7480
70/30	XGBoos t	0.8134	0.3603	0.7538	0.8214	0.4876	0.7876	0.8610	0.4890	0.4538	4.5550
80/20	Logistic Regression	0.7582	0.2935	0.7514	0.7590	0.4221	0.7552	0.8355	0.4622	0.4624	4.6306
80/20	Decision Tree	0.7874	0.3214	0.7283	0.7952	0.4460	0.7618	0.8234	0.4041	0.4566	4.5727
80/20	Random Forest	0.8560	0.4191	0.5838	0.8922	0.4879	0.7380	0.8403	0.4598	0.4509	4.5148
80/20	XGBoos t	0.8118	0.3539	0.7283	0.8229	0.4764	0.7756	0.8416	0.4668	0.4393	4.3990
95/5	Logistic Regression	0.7527	0.2818	0.7209	0.7569	0.4052	0.7389	0.7894	0.3823	0.3023	3.0904
95/5	Decision Tree	0.7962	0.3000	0.5581	0.8277	0.3902	0.6929	0.7551	0.2853	0.2791	2.8527
95/5	Random Forest	0.8315	0.3273	0.4186	0.8862	0.3673	0.6524	0.8030	0.4271	0.3023	3.0904
95/5	XGBoos t	0.7989	0.3176	0.6279	0.8215	0.4219	0.7247	0.8093	0.4170	0.3256	3.3282

Predictive Analytics for Healthcare Cost Reduction

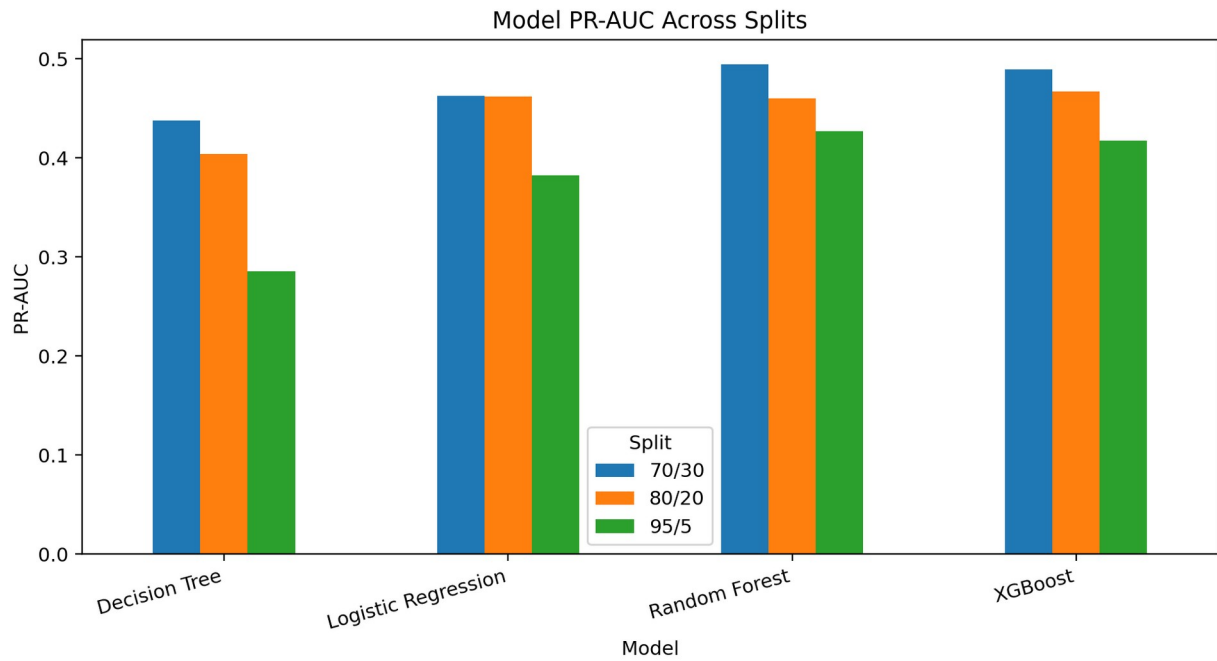


Figure 5. Precision-recall performance across train-test partitions.

Model Evaluation

Random Forest produced the highest average PR-AUC across the three development partitions and was selected for final validation. On the sealed validation set of 818 individuals, the model achieved accuracy 0.874, precision 0.472, recall 0.604, specificity 0.910, F1-score 0.530, ROC-AUC 0.817, and PR-AUC 0.443. The highest predicted-risk decile captured 42.7% of actual high-cost patients, producing lift of 4.31 relative to untargeted selection.

Table 8. Sealed Final Validation Results

Selecte d_Model	Valida tion_N	Accura cy	Precisi on	Recall	Specifi city	F1	Balanc ed_Acc uracy	ROC_ AUC	PR_A UC	Brier	Top_D ecile_C apture	Top_D ecile_L ift
Rando m Forest	818	0.8741	0.4715	0.6042	0.9100	0.5297	0.7571	0.8171	0.4429	0.1037	0.4271	4.3130

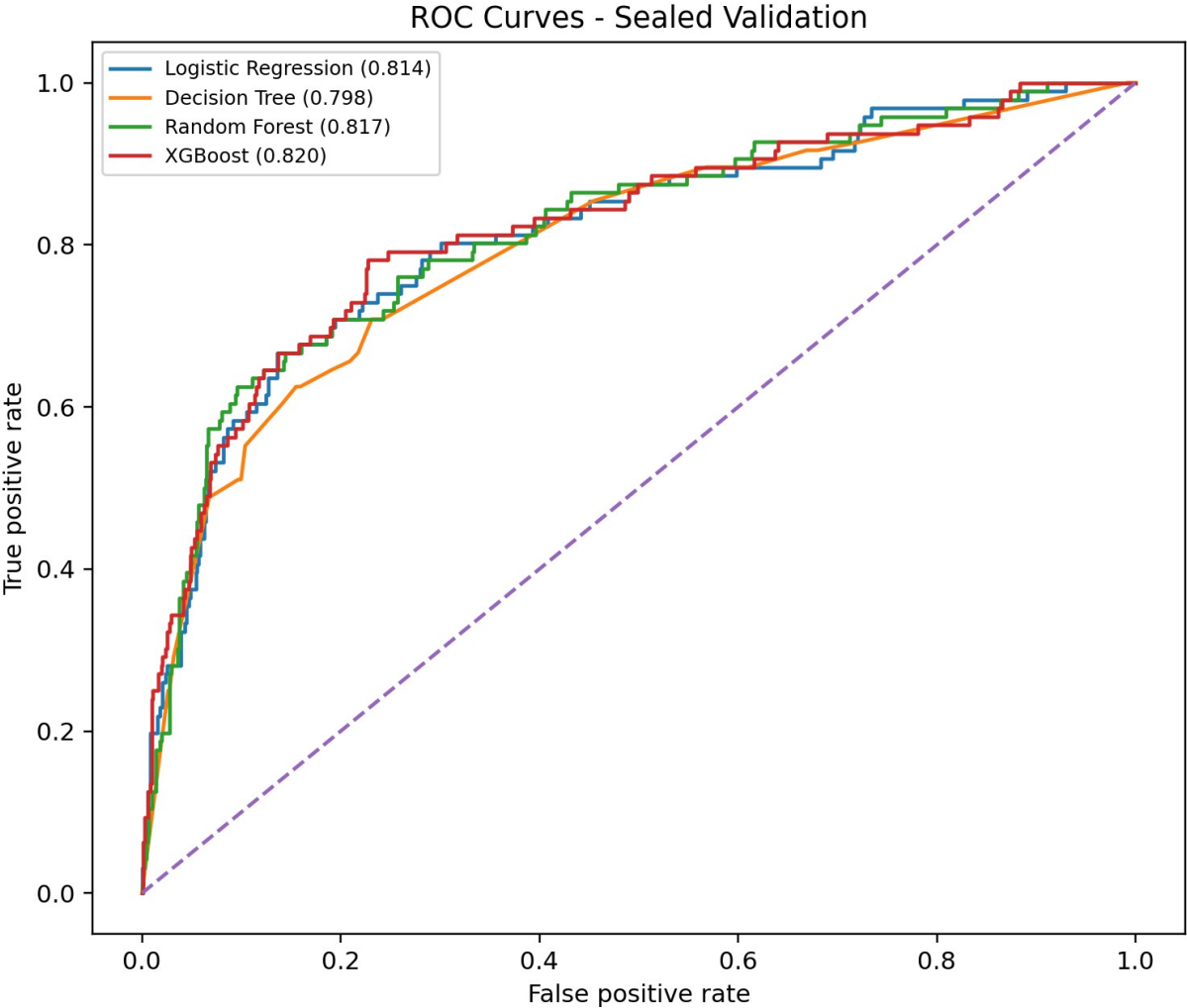


Figure 6. ROC curves on the sealed validation set.

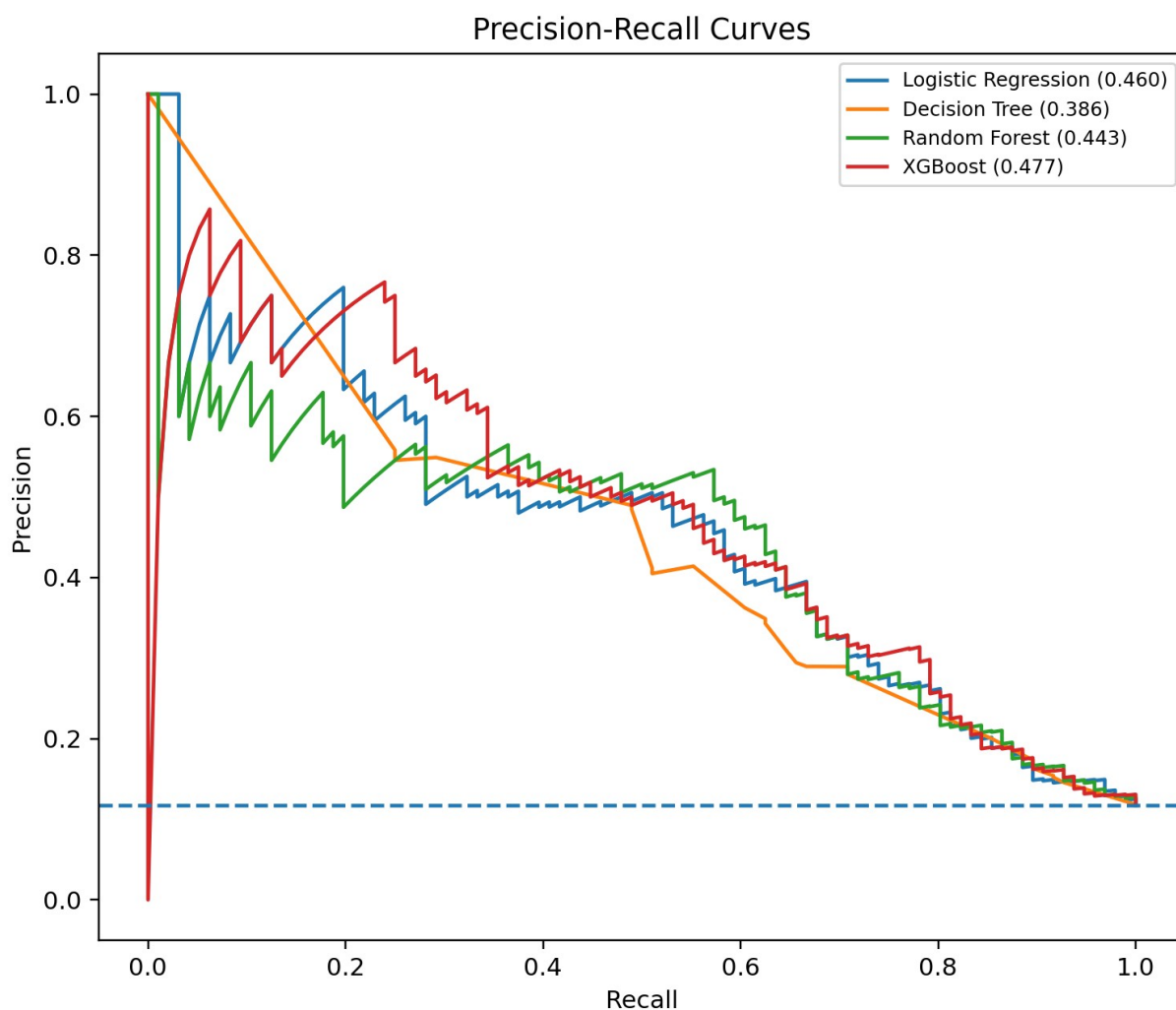


Figure 7. Precision-recall curves on the sealed validation set.

Model Behavior and Interpretation

Table 9. Permutation Feature Importance

Feature	Importance_Mean	Importance_SD
log_prior_expenditure	0.1396	0.0342
prescription_count	0.0486	0.0299
general_health	0.0202	0.0057
walking_limitation	0.0199	0.0043
age	0.0190	0.0105
log_family_income	0.0139	0.0038

Predictive Analytics for Healthcare Cost Reduction

Feature	Importance_Mean	Importance_SD
arthritis	0.0115	0.0071
employed	0.0075	0.0035
cancer	0.0072	0.0021
chronic_burden	0.0064	0.0032
home_health_days	0.0053	0.0021
high_cholesterol	0.0035	0.0014
marital_status	0.0029	0.0021
stroke	0.0014	0.0009
dental_visits	0.0014	0.0045

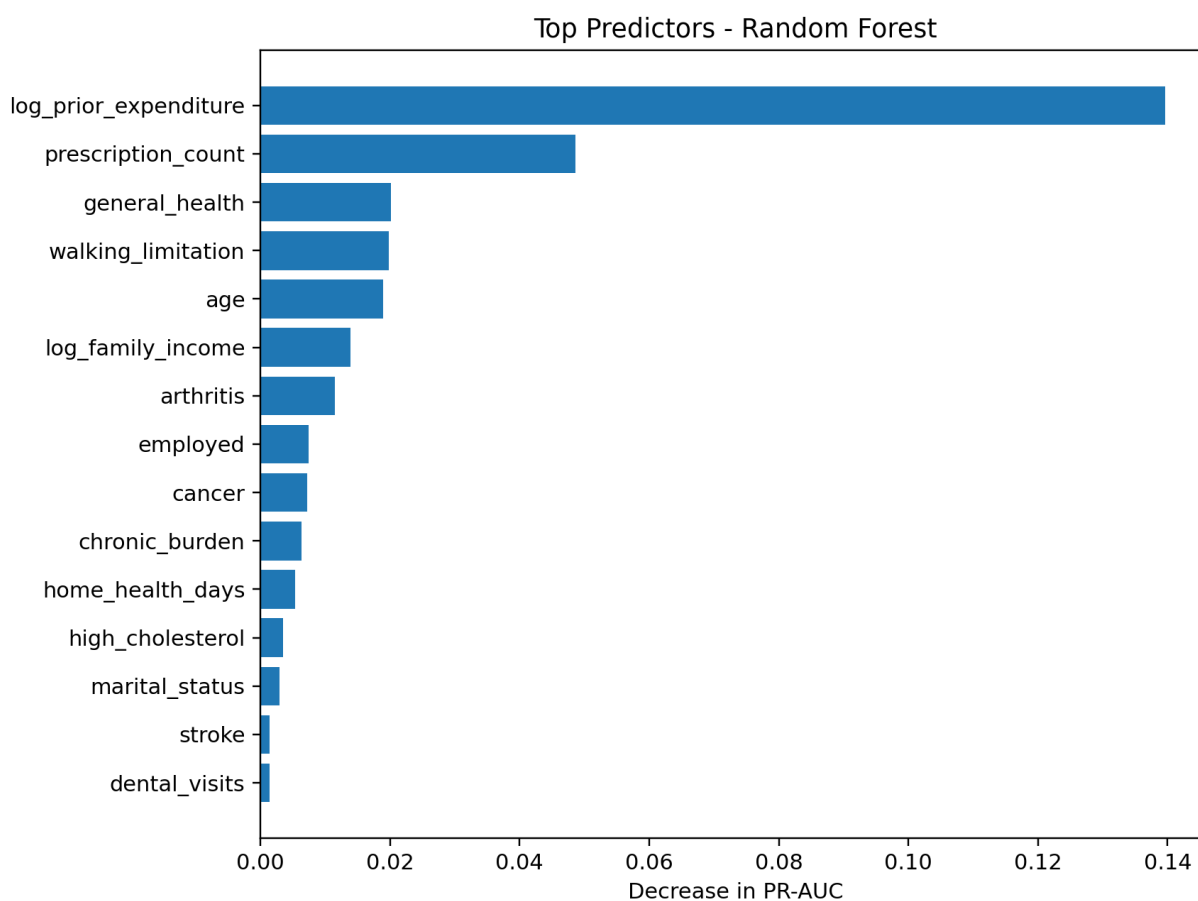


Figure 8. Top predictors of future high-cost status.

Predictive Analytics for Healthcare Cost Reduction

Prior-year expenditure was the dominant predictor, followed by prescription volume, general health, walking limitation, age, family income, arthritis, employment, cancer, chronic-condition burden, and home-health use. These findings are operationally plausible: prior resource use captures existing care intensity, while health status and limitations help distinguish persistent or emerging need. Because permutation importance measures predictive contribution rather than causality, these variables should not be interpreted as intervention effects.

Calibration and Risk Segmentation

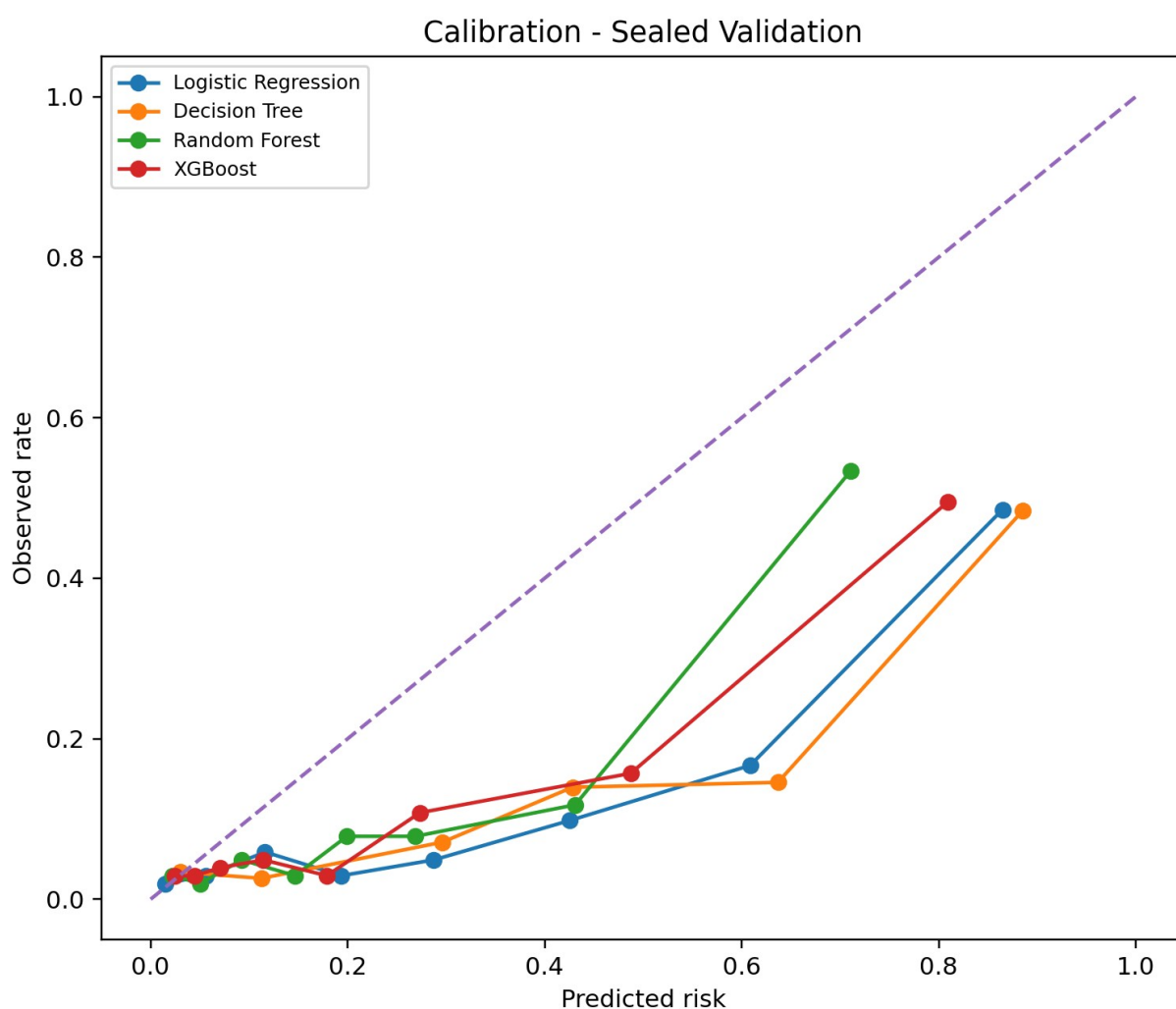


Figure 9. Calibration curves on sealed validation data.

Table 10. Risk Segment Performance

Predictive Analytics for Healthcare Cost Reduction

Risk_Segment	Patients	Observed_High_Cost_Rate	Mean_Predicted_Risk
Very Low	164	0.0244	0.0304
Low	163	0.0429	0.0883
Moderate	164	0.0488	0.1728
High	163	0.0920	0.2852
Very High	164	0.3780	0.6253

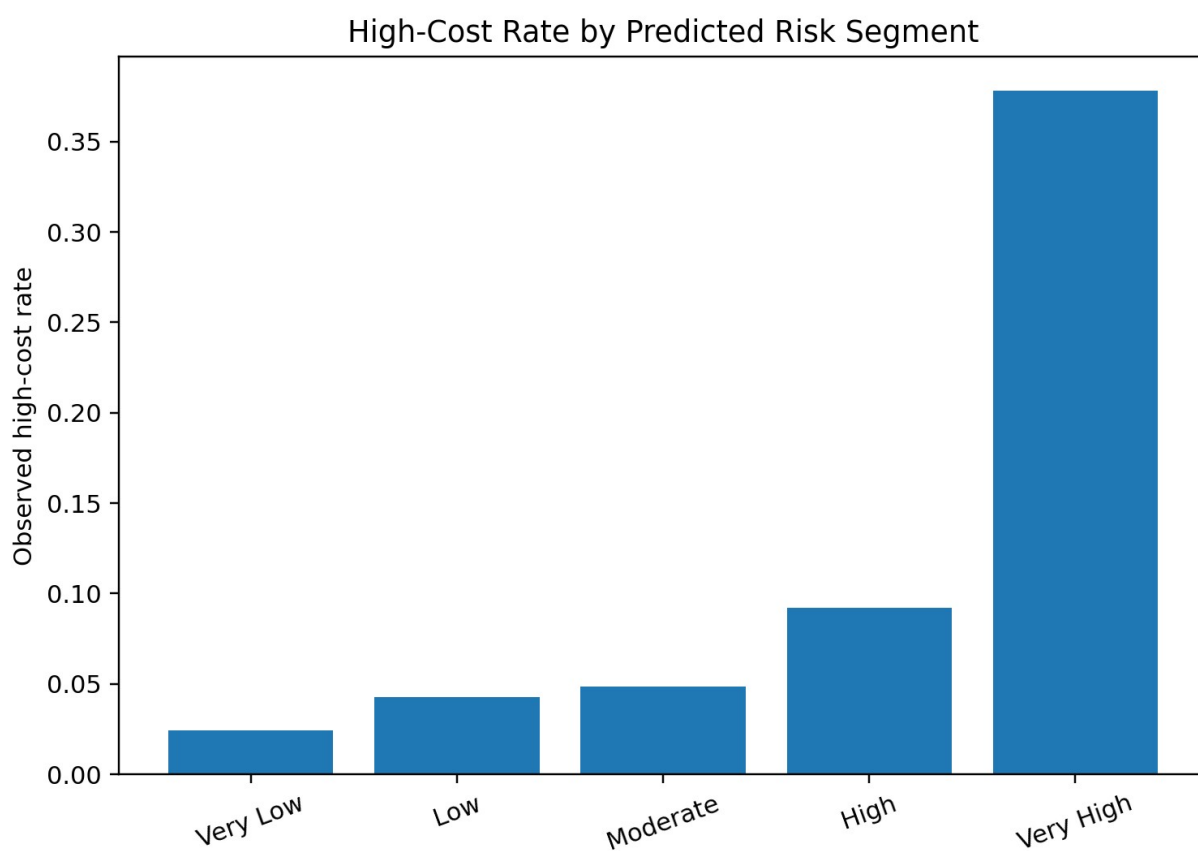


Figure 10. Observed high-cost rate by predicted risk segment.

The observed high-cost rate rose from 2.4% in the very-low-risk group to 37.8% in the very-high-risk group. This monotonic gradient indicates that the score provides useful risk stratification even when a single classification threshold is not optimal for every organization.

Evaluation of Risk

The largest analytic risks are false negatives, false positives, model drift, data-quality variation, and inequitable performance across demographic groups. False negatives may deny preventive resources to people who later incur high costs. False positives may consume care-management capacity without sufficient benefit. Operational deployment should therefore use tiered risk categories, capacity-aware thresholds, and human review rather than a rigid automated decision.

Scenario-Based Stress Testing

Table 11. Synthetic Scenario Stress Test

Scenario	Predicted_High_Cost_Probability
Low-risk adult	0.1048
Moderate chronic risk	0.3376
High utilization risk	0.8176

The stress test produced the expected ordering: the low-risk adult received the lowest predicted probability, the moderate chronic-risk case received an intermediate probability, and the high-utilization case received the highest probability. This does not prove external validity, but it confirms that the model responds directionally to clinically and financially plausible risk patterns.

External Model Verification and Calibration

Literature Review

Prior research supports the feasibility of prospective high-cost prediction while also documenting substantial uncertainty. Langenberger et al. compared machine learning methods for future high-cost patients and emphasized that model performance depends on the prediction horizon, cost definition, and available claims history. Luo et al. found that machine learning could identify potential high-cost patients and highlight influential predictors. Osawa et al. reported that combining screening and claims

Predictive Analytics for Healthcare Cost Reduction

information produced useful discrimination for high-need, high-cost status. Earlier work by Powers et al. demonstrated that prior cost and morbidity information are important for prospective total-cost modeling. The present findings are consistent with this literature because prior expenditure, utilization, health status, and functional limitation emerged as leading predictors, while Random Forest offered stronger ranking performance than simpler alternatives.

External Benchmarks

AHRQ expenditure-concentration reports provide an external benchmark for the target definition. In 2022, the highest-spending 5% of the U.S. civilian noninstitutionalized population accounted for approximately half of healthcare expenditures. Historical MEPS work also shows incomplete year-to-year persistence in the top spending decile, which explains why prospective identification is difficult and why longitudinal modeling adds value.

Calibration

Calibration curves showed that discrimination and probability accuracy are separate concerns. Before operational deployment, predicted probabilities should be recalibrated using the organization's own population, cost definition, intervention capacity, and outcome horizon. Calibration should be reviewed by subgroup and repeated after meaningful changes in coding, reimbursement, utilization, or patient mix.

Model Deployment and Model Life Cycle

Deployment Strategy

1. Data intake and validation: collect prior-year claims, utilization, condition, functional, and socioeconomic information.
2. Batch risk scoring: calculate predicted high-cost probability on a defined schedule.
3. Clinical and operational review: confirm eligibility, remove clearly non-actionable cases, and identify modifiable needs.

Predictive Analytics for Healthcare Cost Reduction

4. Tiered intervention: assign outreach intensity based on predicted risk, expected benefit, and available capacity.
5. Outcome tracking: monitor engagement, emergency utilization, admissions, total costs, patient experience, and equity.
6. Model monitoring: review drift, calibration, false-negative rates, subgroup performance, and threshold effectiveness.
7. Periodic retraining: update the model using new outcomes and revised operational objectives.

Cost-Benefit Framework

Table 12. Illustrative Cost-Benefit Framework

Component	Illustrative Value	Interpretation
Population screened	100,000.00	Illustrative health-system population
Patients targeted	10,000.00	Top predicted-risk decile
Intervention cost per targeted patient	500.00	Care management, navigation, or monitoring
Program investment	5,000,000.00	$10,000 \times \$500$
Average observed high-cost expenditure	48,895.00	Observed HC-252 analytic sample
Required avoided cost for break-even	5,000,000.00	Before administrative overhead and timing adjustments

The cost-benefit table is a transparent scenario, not an estimate of causal savings. A real business case must use local intervention costs, expected effect sizes from validated programs, time to benefit, engagement rates, reimbursement arrangements, and implementation overhead. The predictive model identifies risk; it does not establish that a specific intervention will prevent the predicted expenditure.

Milestones, Training, Cost/Benefit, and Risks

Table 13. Implementation Roadmap

Predictive Analytics for Healthcare Cost Reduction

Milestone	Timing	Key Output
1. Governance and data assessment	Months 1-2	Define outcome, ownership, privacy, fairness, and success criteria
2. Retrospective validation	Months 2-4	Recreate model on local historical data
3. Silent pilot	Months 4-6	Generate scores without changing care; evaluate calibration
4. Controlled intervention pilot	Months 6-12	Test targeted workflow and measure engagement and utilization
5. Scaled deployment	Year 2	Expand with monitored thresholds and retraining
6. Continuous improvement	Ongoing	Drift, equity, safety, and ROI reviews

Recommendations

Recommendations for Practice

- Use the model as a prioritization tool paired with clinical and care-management review.
- Select operating thresholds based on intervention capacity and the relative cost of false negatives and false positives.
- Prioritize PR-AUC, recall, top-decile capture, calibration, and subgroup performance rather than accuracy alone.
- Validate the model on local longitudinal data before any patient-level use.
- Link predictions to evidence-based interventions and separately evaluate whether those interventions improve outcomes and costs.
- Implement governance for privacy, bias, explainability, drift, security, and patient communication.

Recommendations for Future Research

Future work should evaluate multi-year prediction horizons, richer condition histories, pharmacy classes, social determinants, geographic access, provider continuity, and temporal utilization trajectories.

Alternative definitions such as top 5%, preventable hospitalization costs, or cost bloom status should be

tested. Fairness-aware thresholding, survival models, recurrent neural networks, and causal evaluation of targeted interventions may further strengthen the research.

Conclusions

This study developed a prospective high-cost patient prediction framework using nationally oriented longitudinal MEPS data. The final Random Forest model achieved ROC-AUC 0.817 and PR-AUC 0.443 on sealed validation data, captured 42.7% of high-cost patients in the highest predicted-risk decile, and produced lift of 4.31. Prior expenditure, medication volume, health status, mobility limitation, age, and chronic burden were the strongest predictive contributors. The results support rejection of the primary null hypothesis because baseline information contained meaningful predictive signal. The findings also support the use of nonlinear methods for ranking future cost risk, although performance varied across splits and was not sufficient to justify autonomous decisions. A carefully governed implementation could help healthcare organizations focus preventive resources more effectively, but financial benefit depends on local validation and the effectiveness of the interventions connected to the model.

References

Agency for Healthcare Research and Quality. (2025). MEPS HC-252: Panel 27 longitudinal data file.

https://meps.ahrq.gov/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-252

Agency for Healthcare Research and Quality. (2025). MEPS HC-252 Panel 27 longitudinal data public use file documentation. https://meps.ahrq.gov/data_stats/download_data/pufs/h252/h252doc.pdf

Agency for Healthcare Research and Quality. (2024). Concentration of healthcare expenditures and selected characteristics of high spenders, U.S. civilian noninstitutionalized population, 2022.

https://meps.ahrq.gov/data_files/publications/st560/stat560.shtml

Centers for Medicare & Medicaid Services. (2026). National health expenditure data.

<https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data>

Predictive Analytics for Healthcare Cost Reduction

Doupe, P., Faghmous, J., & Basu, S. (2019). Machine learning for health services researchers. *Value in Health*, 22(7), 808-815.

Langenberger, B., et al. (2023). The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data. *PLOS ONE*.

Luo, L., et al. (2020). Using machine learning approaches to predict high-cost patients and identify potential predictors. *BMC Health Services Research*.

Osawa, I., et al. (2020). Machine-learning-based prediction models for high-need high-cost patients using nationwide clinical and claims data. *BMJ Open*.

Powers, C. A., Meyer, C. M., Roebuck, M. C., & Vaziri, B. (2005). Predictive modeling of total healthcare costs using pharmacy claims data. *Medical Care*.

Tamang, S., et al. (2017). Predicting patient cost blooms in Denmark: A longitudinal population study. *BMJ Open*.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, 35(3), 553-572.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*.

Appendix A: Selected Variable Dictionary

Table A1. Core Variables

Variable	Description	Type	Role
TOTEXPY2	2023 total healthcare expenditure	Numeric	Outcome source
HIGH_COST_Y2	Weighted top-decile 2023 expenditure indicator	Binary	Target
TOTEXPY1	2022 total healthcare expenditure	Numeric	Predictor

Predictive Analytics for Healthcare Cost Reduction

Variable	Description	Type	Role
AGEY1X	Age in 2022	Numeric	Predictor
SEX	Sex	Categorical	Predictor
RACETHX	Race/ethnicity category	Categorical	Predictor
POVCATY1	2022 poverty category	Ordinal	Predictor
INSCOVY1	2022 insurance coverage	Categorical	Predictor
RTHLTH3	Self-reported general health	Ordinal	Predictor
MNHLTH3	Self-reported mental health	Ordinal	Predictor
ERTOTY1	2022 emergency visits	Count	Predictor
IPDISY1	2022 inpatient discharges	Count	Predictor
RXTOTY1	2022 prescription events	Count	Predictor
LONGWT	Longitudinal survey weight	Numeric	Estimation weight

Appendix B: Reproducibility and Deliverables

- Original data file: h252.xlsx
- Selected raw variables: h252_selected.csv
- Clean analytic dataset: analytic_dataset.csv
- Model performance: model_performance_all_splits.csv
- Sealed validation results: sealed_validation_metrics.csv
- Predictions: model_predictions_all_splits.csv and sealed_validation_predictions.csv
- Feature importance, risk segmentation, stress tests, and publication-quality figures
- Excel results workbook and full analysis package