

Optimizing Care Management Targeting: A Machine Learning Comparison of Regression, Classification, and Ranking Approaches for Identifying Future High-Cost Patients

Chunhui Mo

Chunhui.Mo@yahoo.com

Independent Researcher <https://orcid.org/0009-0007-8560-6552>

Research Article

Keywords: healthcare expenditure prediction, machine learning, care management, population health management, top-k targeting, MEPS, neural network, classification, regression, ranking

Posted Date: June 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10008979/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

DRAFT

**Optimizing Care Management Targeting: A Machine Learning
Comparison of Regression, Classification, and Ranking Approaches
for Identifying Future High-Cost Patients**

Chunhui Mo, Ph.D.

Independent Researcher

Email: Chunhui.Mo@yahoo.com

ABSTRACT

Effective care management programs depend on accurately identifying patients who are likely to account for a disproportionate share of future healthcare expenditures. Unlike traditional risk-adjustment models, which emphasize population-level predictive accuracy, care management targeting is fundamentally a top-k decision problem focused on maximizing expenditure capture. We used 2000-2023 MEPS data to compare regression-, classification-, and learning-to-rank machine learning approaches for predicting next-year healthcare expenditures. Primary outcomes included precision-at-k, cost capture-at-k, weighted cost capture-at-k, and AUC. Models were benchmarked against a persistence-based floor baseline using prior-year expenditures and an oracle upper bound representing perfect future information. Classification-based models consistently achieved the strongest operational targeting performance. Neural network and XGBoost classifiers outperformed regression and ranking approaches in identifying future high-cost individuals and maximizing expenditure capture. Differences in cost capture were statistically significant across yearly evaluations. More complex modeling extensions provided little additional benefit. Prior-year utilization and expenditure variables contributed substantially to predictive accuracy; however, models retained approximately 80–90% of full-model performance after removal of these features, indicating meaningful predictive signal from sociodemographic characteristics, chronic disease burden, functional limitations, and self-reported health status. For care management programs operating under fixed enrollment constraints, classification-based prediction approaches appear better aligned with operational objectives than regression or ranking. Additional performance gains are more likely to arise from richer clinical and longitudinal data than from increased model complexity alone. Future AI-

enabled care management systems may further improve resource allocation by integrating cost-risk prediction with estimates of patient engagement and intervention impactability.

KEYWORDS : healthcare expenditure prediction, machine learning, care management, population health management, top-k targeting, MEPS, neural network, classification, regression, ranking

INTRODUCTION

Healthcare expenditures in the United States are highly concentrated, with a relatively small proportion of individuals accounting for a disproportionately large share of total spending¹.

Accurate prospective identification of high-risk, high-cost individuals is therefore important for healthcare policy, actuarial forecasting, payer resource allocation, and targeted care management interventions. Predictive modeling of future healthcare expenditures has consequently become an important research area within health services research, medical informatics, and population health management².

Historically, healthcare expenditure prediction models relied primarily on traditional statistical approaches including generalized linear models, logistic regression, and risk-adjustment frameworks such as the Medicare Hierarchical Condition Category (HCC) system. These approaches were designed largely for actuarial calibration and population-level reimbursement adjustment rather than identification of small subsets of patients for intensive intervention programs. Although actuarial risk scores remain highly valuable for payment adjustment and broad population stratification, they are not necessarily optimized for maximizing operational care-management outcomes such as concentration of future expenditures within a limited intervention cohort³.

This distinction has important operational implications because real-world care management programs face substantial resource constraints. Traditional nurse-led case management programs often target only a small fraction of covered populations due to staffing and operational costs. In practice, many commercial programs focus on approximately the top 1–5% highest-risk members, while Medicare-focused programs may target somewhat larger cohorts⁴. Under these

constraints, predictive performance within the extreme upper tail of the expenditure distribution becomes substantially more important than overall calibration across the full population. Consequently, metrics such as precision-at-k and cost capture may better reflect operational value than traditional regression loss metrics alone⁵.

Cost capture—the proportion of total healthcare expenditures accounted for by the predicted high-risk cohort—is particularly important because it directly influences the potential return on investment of care-management interventions. A model that more effectively concentrates future expenditures within a manageable subset of patients can improve targeting efficiency even if its overall regression accuracy is only modestly improved. Despite its practical importance, relatively few studies have systematically compared modeling paradigms specifically from the perspective of top-k expenditure capture³.

Healthcare expenditure prediction is also statistically challenging because medical spending distributions are highly skewed. A small number of catastrophic-cost observations may account for a large proportion of aggregate expenditures, making conventional regression objectives difficult to optimize effectively. Prior studies have shown that standard linear models often struggle with the nonlinear relationships and extreme-value behavior present in healthcare claims and expenditure data^{6,7}.

Recent advances in machine learning have motivated renewed interest in healthcare cost prediction using tree-based ensembles, neural networks, and ranking-based methods. Multiple studies have demonstrated that gradient boosting and neural network approaches can outperform conventional regression models^{3,8}. A recent systematic review reported that machine learning and deep learning models commonly achieve AUC values up to approximately 0.78 for high-risk

classification tasks, with performance gains particularly evident in longitudinal and sequence-rich healthcare datasets⁹. Prior expenditures and utilization history have consistently emerged as the strongest predictors across studies¹⁰.

Most prior studies, however, have evaluated only one or two predictive paradigms at a time.

Healthcare expenditure prediction can naturally be formulated in several different ways. First, it may be treated as a regression problem in which the objective is direct prediction of future expenditures. Second, it may be framed as a classification problem that estimates the probability of belonging to a future high-cost subgroup, such as the top 10% of spenders. Third, it may be approached as a ranking problem in which the objective is to order individuals by relative future expenditure risk. Each formulation implies different optimization objectives and different tradeoffs. For example, regression models may be sensitive to extreme outliers, while classification approaches may penalize misclassification uniformly regardless of the magnitude of missed expenditures. Ranking approaches such as NDCG or pairwise loss may better align with operational targeting objectives because they directly optimize the prioritization of individuals most likely to fall within a high-risk cohort; however, improvements in ranking metrics do not necessarily translate into proportional gains in total expenditures captured^{11–15}.

The Medical Expenditure Panel Survey (MEPS), administered by the Agency for Healthcare Research and Quality (AHRQ), provides a uniquely valuable public dataset for evaluating healthcare expenditure prediction models. MEPS contains nationally representative longitudinal data on demographics, healthcare utilization, insurance coverage, medical conditions, and expenditures, with linkage across consecutive survey years. Importantly, MEPS expenditure data are not synthetic estimates but are constructed using household interviews supplemented by

medical provider and pharmacy validation components, making the dataset widely used for expenditure forecasting, policy analysis, and healthcare utilization studies^{16–18}.

Using MEPS data from 2000–2023, this study systematically compares a range of baseline predictive modeling approaches spanning regression, classification, and ranking paradigms: linear regression, XGBoost regression, neural network regression, logistic regression classification, XGBoost classification, neural network classification, XGBoost ranking with NDCG optimization, and neural network pairwise ranking. Model performance is evaluated using operationally relevant metrics including precision, cost capture, weighted cost capture, and AUC. We additionally benchmark all models against two reference standards: a persistence-based “floor” benchmark using prior-year expenditures and an “oracle” benchmark representing perfect future information.

Beyond predictive performance, we also evaluate model interpretability using SHapley Additive exPlanations (SHAP) for the neural network classification model. SHAP has emerged as one of the most widely used approaches for interpretable machine learning in healthcare because it provides theoretically grounded estimates of feature-level contributions to model predictions⁸. Interpretable modeling is particularly important in healthcare applications where transparency and clinical plausibility influence trust, deployment, and operational adoption⁹.

METHODS

The study used pooled longitudinal Medical Expenditure Panel Survey (MEPS) data spanning 2000–2023. Individual-level observations were extracted using the MEPS variable builder and merged into a unified analytic dataset.

The initial pooled dataset contained 772,800 person-year observations. Sequential inclusion criteria were then applied. First, individuals younger than 18 years were excluded (n removed = 209,050). Second, observations with non-positive annual person-level survey weights were excluded (n removed = 24,633). The eligible dataset contained 539,117 adult person-year observations. To construct the longitudinal prediction cohort, records were linked across consecutive years using the MEPS person identifier (MEPSID). For each individual observed in year (t), feature variables were paired with healthcare expenditures observed in year ($t+1$). Only consecutive-year pairs with observed second year total healthcare expenditure data were retained. This process yielded 257,143 longitudinal person-year pairs used for predictive modeling. Annual total healthcare expenditures, both prior-year and subsequent-year, were inflation-adjusted using national per-capita healthcare expenditure estimates obtained from Federal Reserve Economic Data.

Because MEPS panels partially overlap across survey years, some individuals contributed multiple consecutive-year observation pairs. Duplication analysis demonstrated that 235,988 unique MEPSIDs generated 257,143 longitudinal records, with 11,922 individuals contributing more than one observation pair. Most repeated individuals contributed either two or three sequential year-pairs, reflecting the rotating overlapping panel structure of MEPS rather than duplicate records. These repeated observations were retained to preserve the longitudinal

continuity and maximize available training data for rolling-year prediction experiments. We did not explicitly model within-person correlation across repeated observations because repeated individuals represented a small proportion of the analytic cohort (5.05%).

Feature engineering was performed using the longitudinal MEPS panel dataset constructed from consecutive-year person-level observations. The final analytic modeling dataset included observations from 2000–2023 and retained adjusted prior-year expenditures as a feature and adjusted subsequent-year expenditures as the prediction target. Features were grouped into five clinically and conceptually motivated domains: sociodemographic characteristics, healthcare utilization, chronic disease burden, functional and self-reported health status, and insurance coverage.

Sociodemographic features included age, poverty category, educational attainment, sex, race/ethnicity, and geographic region. Age was treated as a continuous variable after conversion to numeric format. Invalid MEPS age codes (e.g., 996) were treated as missing and median-imputed due to minimal missingness. Educational attainment was recoded from detailed MEPS categories into five ordinal levels representing: less than high school, high school/GED, some college/associate degree, bachelor's degree, and graduate education. Missing or noninformative education codes were treated as missing, and an accompanying binary missingness indicator was created. Remaining missing values were median-imputed. Race and ethnicity variables were harmonized into grouped race/ethnicity categories using MEPS race variables with Hispanic ethnicity overriding race classification. Geographic region was encoded using ordinal region categories derived from MEPS census region variables.

Healthcare utilization features were derived exclusively from year (t) information. Features included outpatient office visits, emergency department visits, inpatient hospital discharges, prescription medication counts, and prior-year adjusted total expenditures. A composite utilization intensity index was created by combining office visits, emergency room visits, and hospital discharges into a single weighted measure that assigned greater emphasis to higher-acuity services.

Chronic disease features were aggregated into clinically related condition domains using MEPS condition indicators. Four grouped disease burden indices were created: cardiovascular disease burden, metabolic disease burden, respiratory disease burden, and other chronic disease burden. Within each domain, binary MEPS disease indicators were summed to create condition burden indices representing the number of chronic conditions present within each system category.

Additional missingness-count variables were created for each chronic disease domain to preserve information related to incomplete condition reporting. The cardiovascular disease domain included myocardial infarction, coronary heart disease, angina, stroke, hypertension, and congestive heart failure indicators. Metabolic disease variables included diabetes and hyperlipidemia indicators. Respiratory disease variables included asthma and emphysema indicators. Additional chronic conditions included cancer, arthritis, and attention-deficit disorder indicators.

Functional and health status features included self-reported health status, functional limitation indicators, exercise frequency, loneliness measures, and psychological distress scores. Self-reported health status was modeled as an ordinal variable using the MEPS five-level health scale ranging from excellent to poor health. Functional limitation indicators were converted into

binary variables and supplemented with explicit missingness indicators. Exercise frequency, loneliness measures, and Kessler-6 psychological distress scores were treated as continuous or ordinal variables following numeric conversion. For variables with low levels of missingness, median or mode imputation was used to preserve sample size and maintain modeling stability. Separate binary missingness indicators were retained for variables where missingness itself could potentially convey predictive information.

Insurance features included indicators for private insurance coverage, Medicaid/CHIP coverage, Medicare coverage, and uninsured status. Insurance variables were standardized into binary indicator format.

MEPS-specific noninformative response codes (e.g., “don’t know,” “refused,” “not ascertained,” and “not applicable”) were systematically recoded as missing values before feature construction. Continuous and ordinal variables with low missingness were imputed using median values, while selected categorical variables were imputed using modal categories or clinically conservative defaults. Missingness indicator variables were retained for several predictor domains to preserve potential information associated with incomplete responses. The final feature set consisted of 35 engineered features together with survey year.

The primary prediction task focused on identifying individuals in the top 10% of future healthcare expenditures ($k = 10\%$). To evaluate threshold robustness, additional analyses were conducted at ($k = 5\%$) and ($k = 30\%$). Binary classification labels were generated within each year using percentile rankings of adjusted expenditures for classifiers.

We evaluated a range of baseline modeling approaches spanning regression, classification, and ranking paradigms. Regression models included a linear regressor, an XGBoost regressor, and a feedforward neural network regressor trained using Huber loss on log-transformed expenditures. Classification models included a logistic regression classifier, an XGBoost classifier and a neural network classifier trained using binary cross-entropy (BCE) loss. Ranking approaches included an XGBoost learning-to-rank ranker optimized using an NDCG objective and a neural network pairwise ranker trained using pairwise binary cross-entropy on sampled observation pairs. The neural network models used a multilayer perceptron architecture consisting of fully connected hidden layers with ReLU activations, batch normalization, and dropout regularization. Optimization was performed using the Adam optimizer.

We also evaluated a few extended model specifications to assess whether alternative training objectives or multi-stage learning strategies could further improve predictive performance. For neural network models, we considered classifiers trained with focal loss to address class imbalance and top-k weighting to emphasize performance in the upper tail of the expenditure distribution. We also evaluated a two-stage approach, combining an initial classification model with a subsequent ranking model applied to high-probability observations. In the XGBoost-based pipeline, the second-stage model used a learning-to-rank objective optimized for NDCG, while the neural network-based pipeline used a pairwise ranking loss. All models were trained and evaluated under the same data-splitting and testing framework as the baseline specifications.

To evaluate predictive performance in a temporally realistic setting, we implemented a rolling-origin evaluation design. Models were trained using an expanding window of historical data and tested on a single out-of-sample year. Specifically, for each evaluation iteration, all data from

2000 up to year t were used for training, and model performance was evaluated on year $t+1$, with t ranging from 2009 to 2021. This procedure resulted in 13 sequential out-of-sample test sets, ensuring that all performance metrics reflect genuine forward-looking prediction. All models, including baseline and extended specifications, were estimated and evaluated consistently within this rolling projection framework.

Due to the limited number of available time periods and the presence of potential structural breaks in the data (including MEPS questionnaire redesign around the 2017–2018 transition¹⁹ and the COVID-19 pandemic period), we did not employ a separate validation set or early stopping procedure. Instead, all model training and evaluation were conducted within the rolling-origin framework using expanding training windows. The objective of this study was comparative evaluation of modeling paradigms rather than optimization of individual algorithms. As a result, we also did not perform dataset-specific hyperparameter tuning. Model specifications were held fixed across all rolling evaluation folds to ensure comparability and avoid additional temporal partitioning of the data. This design prioritizes consistency and out-of-sample evaluation integrity over fold-specific parameter optimization.

Three feature configurations were evaluated: a full feature set, a reduced set excluding prior-year expenditures (“No Lagged Cost”), and a further reduced specification excluding all utilization-related features (“No Lagged Utilization”). These scenarios were designed to assess the dependence of predictive performance on historical spending and utilization information.

Model performance was evaluated using area under the receiver operating characteristic curve (AUC), precision at k , cost capture at k , and weighted cost capture at k . Precision measures the proportion of individuals correctly identified within the top $k\%$ of predicted high-cost patients.

For AUC, regression predictions were converted to ranked risk scores and evaluated against binary top-10% membership labels. Cost capture was defined as the proportion of total realized expenditures in the prediction year attributable to the top k% of patients based on predicted risk, calculated as the sum of their observed expenditures divided by the total expenditures of the full population. Survey weights were not incorporated during model training because the primary objective was individual-level prediction rather than population parameter estimation. Weighted cost capture was calculated using the same approach, except that expenditures were weighted by the MEPS final annual person-level weights (PERWT) to produce nationally representative estimates. In addition to model-based predictions, two benchmark comparators were included: (1) a floor benchmark computed using prior-year expenditures alone and (2) an oracle benchmark computed using the true future expenditures to represent the theoretical upper bound. P-values from pairwise model comparisons are reported descriptively; no adjustment for multiple comparisons was applied given the exploratory nature of the comparisons.

To assess model interpretability, SHAP (SHapley Additive exPlanations) analyses were performed for the neural network classification model. SHAP values were computed using GradientExplainer on held-out test data, and global feature importance was summarized using the mean absolute SHAP value across observations.

RESULTS

We evaluated predictive performance of a range of baseline models - three regression models (linear regressor, gradient boost (XGB) regressor and neural network (NN) regressor), three classification models (logistic regression classifier, XGB classifier and NN classifier) and two ranking models (XGB ranker and NN ranker), using 13 rolling-year evaluations (2010–2022) at $k = 10\%$, where models are evaluated on their ability to identify the top 10% highest-cost patients in the following year (Table 1). The floor baseline is computed using only prior-year expenditure, and the oracle prediction upper bound is computed using realized total expenditure in the evaluation year. Performance is reported using AUC, precision, cost capture, and weighted cost capture. All models achieve better performance than the floor benchmark but the gap to the oracle benchmark remains substantial. Weighted cost capture closely tracks standard cost capture across models, indicating that performance gains are consistent across both unweighted and weighted formulations using MEPS final basic annual weight.

Across all metrics, classifiers (neural network or XGBoost) achieve strong performance in both discrimination (AUC) and cost capture, while regressors show intermediate performance between classifiers and rankers (Figure 1). Neural network and XGBoost models performed better than linear models (linear regressor, logistic regression classifier). Pairwise Wilcoxon signed-rank tests over matched yearly observations indicate statistically significant differences in cost capture among modeling approaches and relative to the floor benchmark (Figure 2). Each cell corresponds to a pairwise comparison of model performance across identical evaluation years, ensuring a paired design. Paired t-tests yielded similar conclusions (results not shown).

We also evaluated a set of extended modeling strategies, including two-stage architectures (combining a classification stage with a subsequent ranking model) and neural network classifiers trained with alternative loss functions (focal loss and top-k weighting). As shown in Table 2, these extensions produced only modest and inconsistent changes in cost capture performance. Importantly, none of the alternative specifications consistently outperformed the standard neural network classifier trained with BCE, suggesting limited additional gains from more complex model specifications in this setting.

Temporal stability in model performance is tracked annually across the 2010–2022 evaluation window (Figure 3a). The floor model provides a stable lower-bound baseline across years, reflecting persistence in expenditure patterns driven by prior-year total costs. The oracle benchmark represents an upper-bound constructed using realized next-year total expenditures and therefore defines the maximum achievable cost capture under perfect foresight. The neural network classifier shows consistent performance improvements over the floor benchmark across most years, with some variation in magnitude over time. The observed performance fluctuations around 2018–2020 coincide with known structural changes in the MEPS survey design period and the onset of the COVID-19 pandemic, both of which may have affected expenditure measurement and utilization patterns¹⁹.

Model performance sensitivity to prediction threshold is evaluated for top 5%, 10%, and 30% highest risk population (Figure 3b). As expected, cost capture increases as the prediction set expands. The neural network model maintains a consistent performance advantage over the floor benchmark across all thresholds, while the oracle benchmark provides an upper bound that remains substantially higher across all settings.

Predictive performance varies systematically across sub-groups with different insurance coverage (Figure 3c). The privately insured population exhibits the strongest separation between the neural network model and the floor baseline, indicating the highest marginal predictive value of the model. The Medicare subgroup exhibited smaller incremental improvements over the floor, consistent with a population that is more clinically homogeneous and whose high-cost dynamics are less sensitive to short-term predictive signals. Medicaid shows the weakest and most variable performance gains, reflecting greater population instability, including higher turnover, intermittent coverage, and more volatile cost trajectories.

The importance of utilization feature variables is evaluated by examining model performance under progressively restricted information environments (Figure 4a). We compare three specifications: Full Features, No Lagged Cost (removing prior-year total healthcare expenditure), and No Lagged Utilization (removing prior-year utilization measures including prescriptions, office visits, emergency room visits, hospitalizations, total utilization indices, and prior-year total healthcare expenditure). As expected, the Full Features specification achieves the highest cost capture relative to the oracle benchmark and outperforms the floor benchmark, confirming that both lagged cost and lagged utilization variables contain important predictive information.

Removing lagged utilization variables results in a moderate reduction in performance, with the neural network retaining approximately ~80% of full-model predictive performance. In contrast, removing only prior-year total expenditure leads to a smaller decline, with performance remaining at approximately ~90% of full-model predictive performance. In both cases, model performance remained below the floor benchmark. Overall, these results suggest that prior utilization and cost history constitute the primary drivers of predictive accuracy. Nevertheless,

even when these signals are partially removed, the model retains a substantial fraction of predictive performance, indicating that other feature variables (e.g., sociodemographic, health status, and chronic disease variables) carry meaningful predictive information.

Feature importance and model interpretability is further evaluated for both individual features and higher-level domains across model specifications. SHAP-based rankings of the top predictive features show a small subset of variables consistently dominates predictive performance across all specifications (Figure 4b). In particular, self-reported health status, age, sex, indicators of functional limitation (anylmt), cardiovascular disease burden (cvd_index), and private insurance coverage (hiprivate) appear among the most influential predictors in all scenarios. A notable pattern is the stability of these core predictors across feature restrictions. Even when prior-year utilization and cost variables are removed, the model continues to rely heavily on demographic structure (age, sex), baseline health status, and chronic disease burden. This suggests that a substantial portion of predictive signal in healthcare expenditure is embedded in relatively stable patient characteristics rather than short-term utilization dynamics alone. At the same time, removing lagged utilization and prior expenditure features shifts the relative importance distribution, increasing reliance on functional status and self-reported health variables, consistent with these variables acting as partial substitutes for prior cost history.

SHAP contributions are aggregated into clinically and conceptually meaningful feature domains, providing a more structured view of model reliance on different categories of information (Figure 4c). Across all specifications, prior utilization and functional and health status domains jointly account for the largest share of predictive signal in the full model, followed by sociodemographic, chronic disease and insurance variables.

Figure 4 indicates that predictive performance is not driven by a large set of weak predictors, but instead by a small and stable subset of clinically interpretable variables. This pattern holds across feature ablations, suggesting that the model does not rely on a single dominant signal source, but rather integrates multiple stable health and demographic dimensions. The consistency of SHAP importance across specifications further supports the robustness of the learned representations, while the domain-level aggregation confirms that medically meaningful structure is preserved even under restricted feature sets.

DISCUSSION

Patient risk stratification represents the first step in the traditional care management workflow, which typically progresses through identification, engagement, activation, and long-term maintenance. Prior literature has shown that structured care management programs can reduce avoidable healthcare utilization and improve outcomes, although their effectiveness depends heavily on accurately identifying high-risk patients and efficiently allocating limited care management resources^{1,4,20}.

AI-based prediction systems have the potential not only to improve targeting accuracy but also to reduce operational costs across the care continuum. Recent work in clinical machine learning and health systems AI suggests that predictive models can improve identification efficiency, but also raise the possibility of workflow redesign where prediction is embedded into automated outreach and triage systems rather than static risk scoring tools. Prior studies have demonstrated that patient-reported health status measures—including functional limitations, self-rated general health, and symptom burden—provide independent predictive value for future healthcare utilization and cost trajectories beyond administrative and clinical risk factors^{21–23}. AI-driven systems could be used to expand initial engagement and collect structured patient and caregiver inputs, which may provide incremental predictive signal beyond traditional clinical encounters. This direction is particularly relevant given constraints in clinical interaction time. Evidence suggests that physicians spend only a limited amount of time on patient interaction per visit, often leaving limited capacity for structured risk assessment or social determinant elicitation²⁴. This reinforces the potential role of digital and AI-assisted data collection in augmenting contemporary care management. Beyond improving prediction of future healthcare expenditures,

AI systems may eventually support a more dynamic and adaptive care management framework by estimating not only future cost risk but also the probability of patient engagement, the likelihood that costs are modifiable through intervention, and the expected impact of specific care management actions on future outcomes. Such capabilities could enable a closed-loop care management paradigm in which patient outreach, engagement, intervention, and outcome measurement continuously inform model refinement and operational decision-making, supporting ongoing learning and performance improvement.

The MEPS survey design introduces several important constraints for predictive modeling. First, the dataset provides only a short longitudinal window for most individuals (typically two-year panels), limiting the ability to capture longer-term expenditure dynamics, chronic disease progression trajectories, or multi-year utilization “momentum” effects. In contrast, administrative claims datasets (e.g., Medicare claims or large commercial insurer databases) provide richer longitudinal depth, enabling modeling approaches that incorporate multi-year history, cumulative burden measures, and trajectory-based features. MEPS is also subject to panel attrition and survey survival bias. Individuals with more severe illness or higher utilization may be differentially likely to drop out or have incomplete follow-up, which can bias longitudinal relationships. This limitation is widely acknowledged in MEPS-based econometric and health services research and must be considered when generalizing predictive performance to real-world payer populations²⁵.

One possible explanation for the superior performance of classification models is that the operational objective itself is threshold-based, although this interpretation remains speculative and was not directly tested. Care management programs ultimately select a finite intervention

cohort, making the prediction task closely aligned with estimating membership in a future high-cost group rather than optimizing global rank ordering across the entire expenditure distribution.

An important finding is that the floor benchmark based solely on prior-year expenditures remains highly competitive, emphasizing the substantial persistence present in healthcare spending trajectories. This result aligns with a long-standing finding in healthcare expenditure modeling literature: cost trajectories exhibit strong persistence over time, with a large share of predictive signal concentrated in lagged utilization and prior spending levels. The strong performance of simple lag-based benchmarks (“floor”) confirms this persistence effect and highlights a key limitation in predictive modeling for healthcare cost: marginal gains from increasingly complex models may be constrained unless additional sources of signal are introduced. In addition, none of the extended modeling strategies tested, including two-stage architectures and neural network classifiers trained with alternative loss functions (focal loss and top-k weighting), outperformed the standard neural network classifier trained with BCE. This pattern reinforces prior findings in healthcare expenditure prediction that lagged utilization and expenditure variables account for a substantial share of predictive signal, often limiting the incremental gains achievable through more complex modeling approaches²⁶. This observation is also consistent with prior machine learning literature in healthcare prediction tasks, which has found that performance gains from increased model complexity often diminish once strong structured tabular features (e.g., prior utilization and cost history) are included^{27,28}.

This study also demonstrates that prior utilization dominates prediction, but substantial predictive signal remains in sociodemographics, health status, and chronic disease burden. It aligns with evidence that substantive improvements in predictive performance are more often

driven by feature enrichment—such as incorporation of social determinants of health, richer longitudinal claims histories, or learned clinical representations—rather than model complexity alone^{29–31}. Future work should focus on integrating richer longitudinal datasets, incorporating patient-reported and social determinants of health, and exploring intervention-aware predictive objectives that directly optimize care management return on investment rather than purely statistical accuracy.

ACKNOWLEDGEMENTS

This study uses data from the Medical Expenditure Panel Survey (MEPS) administered by the Agency for Healthcare Research and Quality (AHRQ). We acknowledge AHRQ for providing publicly available survey data and supporting documentation, including the MEPS variable construction resources that made this analysis possible.

We also gratefully acknowledge the Python open-source ecosystem, including libraries for data processing, machine learning, and visualization, which enabled the full end-to-end implementation of this study within a reproducible and scalable framework.

Finally, we acknowledge the use of ChatGPT (OpenAI) as a collaborative tool in supporting iterative development of data preprocessing, feature engineering, model implementation, and result analysis.

REFERENCES

1. Powers BW, Chaguturu SK, Ferris TG. Optimizing High-Risk Care Management. *JAMA*. 2015;313(8):795-796. doi:10.1001/jama.2014.18171
2. de Ruijter UW, Kaplan ZLR, Bramer WM, et al. Prediction Models for Future High-Need High-Cost Healthcare Use: a Systematic Review. *J Gen Intern Med*. 2022;37(7):1763-1770. doi:10.1007/s11606-021-07333-z
3. Morid MA, Kawamoto K, Ault T, Dorius J, Abdelrahman S. Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. *AMIA Annu Symp Proc*. 2018;2017:1312-1321.
4. DuBard CA, Jackson CT. Active Redesign of a Medicaid Care Management Strategy for Greater Return on Investment: Predicting Impactability. *Popul Health Manag*. 2018;21(2):102-109. doi:10.1089/pop.2017.0122
5. Maisog JM, Li W, Xu Y, et al. Using massive health insurance claims data to predict very high-cost claimants: a machine learning approach. *arXiv*. Preprint posted online December 30, 2019:arXiv:1912.13032. doi:10.48550/arXiv.1912.13032
6. Wu Z, Berkowitz SA, Heagerty PJ, Benkeser D. A two-stage super learner for healthcare expenditures. *Health Serv Outcomes Res Methodol*. 2022;22(4):435-453. doi:10.1007/s10742-022-00275-x
7. Drewe-Boss P, Enders D, Walker J, Ohler U. Deep learning for prediction of population health costs. *BMC Med Inform Decis Mak*. 2022;22(1):32. doi:10.1186/s12911-021-01743-z
8. Orji U, Ukwandu E. Machine Learning For An Explainable Cost Prediction of Medical Insurance. *arXiv*. Preprint posted online November 23, 2023:arXiv:2311.14139. doi:10.48550/arXiv.2311.14139
9. Lee JT, Yeh MHS, Li VCS, et al. Comparing deep learning and classical regression approaches for predicting healthcare expenditure and spending: a systematic review. *J Med Econ*. 2026;29(1):654-671. doi:10.1080/13696998.2026.2630598
10. Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting high cost high need patient expenditures in health care. *Biomed Eng OnLine*. 2018;17(Suppl 1):131. doi:10.1186/s12938-018-0568-3
11. Burges C, Shaked T, Renshaw E, et al. Learning to rank using gradient descent. In: *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*. ACM Press; 2005:89-96. doi:10.1145/1102351.1102363
12. Burges CJC. From RankNet to LambdaRank to LambdaMART: An Overview. *Microsoft Res Tech Rep MSR-TR-2010-82*.
13. Manning WG, Mullahy J. Estimating log models: to transform or not to transform? *J Health Econ*. 2001;20(4):461-494. doi:10.1016/s0167-6296(01)00086-8

14. Bishop CM. *Pattern Recognition and Machine Learning*. Springer; 2006.
15. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.
16. Moeller J, Cohen S. Methodology Report #13: Projecting National Medical Expenditure Survey Data: A Framework for MEPS Projections. Accessed June 3, 2026.
https://meps.ahrq.gov/data_files/publications/mr13/mr13.shtml
17. Bernard D, Cowan C, Selden T, Cai L, Catlin A, Heffler S. Reconciling Medical Expenditure Estimates from the MEPS and NHEA, 2007. *Medicare Medicaid Res Rev*. 2012;2(4):mmrr.002.04.a09. doi:10.5600/mmrr.002.04.a09
18. Wang J, Shi L. Prediction of medical expenditures of diagnosed diabetics and the assessment of its related factors using a random forest model, MEPS 2000–2015. *Int J Qual Health Care*. 2020;32(2):99-112. doi:10.1093/intqhc/mzz135
19. MEPS HC-206G. Accessed June 5, 2026.
https://meps.ahrq.gov/data_stats/download_data/pufs/h206g/h206gdoc.shtml?utm_source=chatgpt.com
20. Berry-Millett R, Bodenheimer TS. Care management of patients with complex health care needs. *Synth Proj Res Synth Rep*. 2009;(19):52372.
21. Boscardin CK, Gonzales R, Bradley KL, Raven MC. Predicting cost of care using self-reported health status data. *BMC Health Serv Res*. 2015;15:406. doi:10.1186/s12913-015-1063-1
22. Barker AR, Joynt Maddox KE, Peters E, Huang K, Politi MC. Predicting Future Utilization Using Self-Reported Health and Health Conditions in a Longitudinal Cohort Study: Implications for Health Insurance Decision Support. *Inq J Med Care Organ Provis Financ*. 2021;58:00469580211064118. doi:10.1177/00469580211064118
23. Perrin NA, Stiefel M, Mosen DM, Bauck A, Shuster E, Dirks EM. Self-reported health and functional status information improves prediction of inpatient admissions and costs. *Am J Manag Care*. 2011;17(12):e472-478.
24. Sinsky C, Colligan L, Li L, et al. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. *Ann Intern Med*. 2016;165(11):753-760. doi:10.7326/M16-0961
25. Cohen SB, Machlin SR. Survey attrition considerations in the Medical Expenditure Panel Survey. *J Econ Soc Meas*. 2000;26(2):83-98. doi:10.3233/JEM-2000-0173
26. Fleishman JA, Cohen JW. Using Information on Clinical Conditions to Predict High-Cost Patients. *Health Serv Res*. 2010;45(2):532-552. doi:10.1111/j.1475-6773.2009.01080.x
27. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Calster BV. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004

28. Vimont A, Leleu H, Durand-Zaleski I. Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in France. *Eur J Health Econ.* 2022;23(2):211-223. doi:10.1007/s10198-021-01363-4
29. Adler NE, Stead WW. Patients in Context — EHR Capture of Social and Behavioral Determinants of Health. *N Engl J Med.* 2015;372(8):698-701. doi:10.1056/NEJMp1413945
30. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. *J Am Med Inform Assoc JAMIA.* 2020;27(11):1764-1773. doi:10.1093/jamia/ocaa143
31. Tan M, Hatef E, Taghipour D, et al. Including Social and Behavioral Determinants in Predictive Models: Trends, Challenges, and Opportunities. *JMIR Med Inform.* 2020;8(9):e18084. doi:10.2196/18084

Model	AUC	Precision	Cost Capture	Weighted Cost Capture
Floor	0.832	0.442	0.380	0.374
Linear Regressor	0.830	0.456	0.386	0.376
XGB Ranker	0.843	0.467	0.397	0.389
Logistic Classifier	0.848	0.471	0.398	0.388
NN Ranker	0.851	0.480	0.401	0.392
XGB Regressor	0.854	0.476	0.404	0.395
NN Regressor	0.851	0.485	0.405	0.397
XGB Classifier	0.858	0.488	0.410	0.400
NN Classifier	0.857	0.490	0.412	0.404
Oracle	1.000	1.000	0.645	0.638

Table 1. Summary statistics across models. Prediction results for all models and a floor baseline using prior-year expenditure, and an oracle upper bound computed using realized total expenditure in the evaluation year are averaged over 13 rolling-year evaluations (2010–2022). Metrics are reported at $k = 10\%$, corresponding to prediction of the top 10% highest-cost patients in the following year.

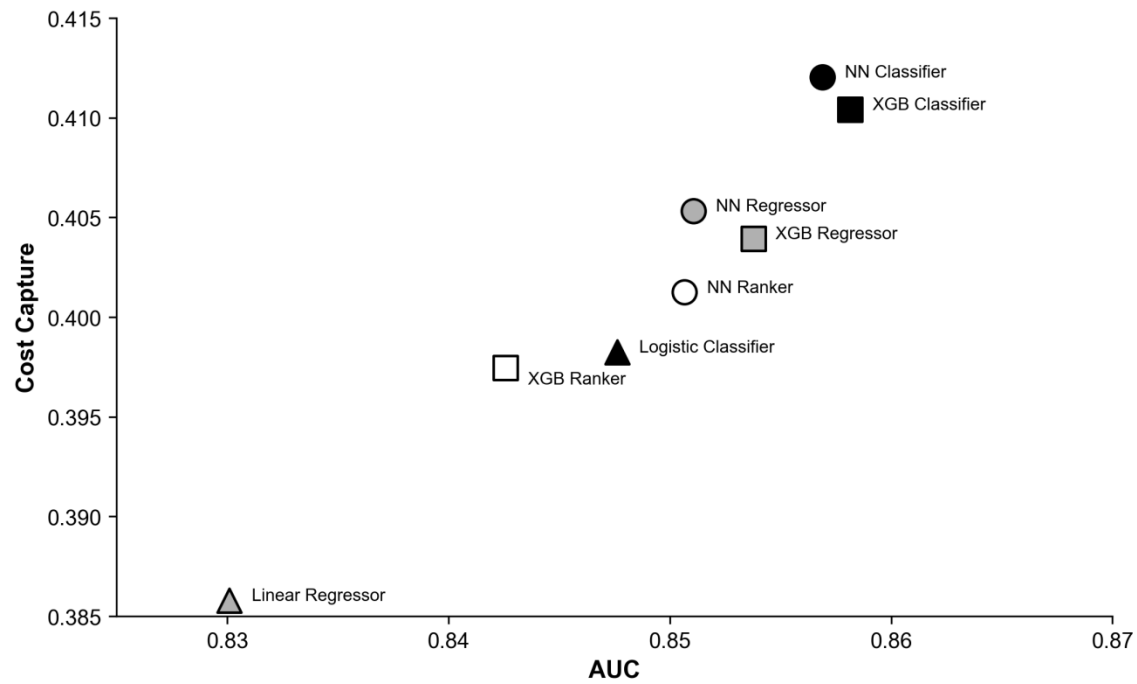


Figure 1. Scatter plot of model performance. Each point represents a model averaged over 13 rolling-year evaluations (2010–2022). The x-axis shows AUC, measuring discrimination performance. The y-axis shows Cost Capture at $k = 10\%$, defined as the proportion of total following year healthcare expenditure captured within the top 10% highest-cost predicted patients. Marker shape denotes model family (circle = neural network, square = XGBoost, triangle = linear model). Marker fill denotes prediction task (black = classifier, gray = regressor, unfilled = ranker).

Floor		0.021	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Linear Regressor	0.021		0.216	0.021	0.946	0.340	0.191	<0.001	<0.001	<0.001
XGB Ranker	<0.001	0.216		0.839	0.455	0.168	0.127	<0.001	<0.001	<0.001
Logistic Classifier	<0.001	0.021	0.839		0.168	0.040	0.013	<0.001	<0.001	<0.001
NN Ranker	<0.001	0.946	0.455	0.168		0.414	<0.001	<0.001	<0.001	<0.001
XGB Regressor	<0.001	0.340	0.168	0.040	0.414		0.414	0.002	<0.001	<0.001
NN Regressor	<0.001	0.191	0.127	0.013	<0.001	0.414		0.005	<0.001	<0.001
XGB Classifier	<0.001	<0.001	<0.001	<0.001	<0.001	0.002	0.005		0.340	<0.001
NN Classifier	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.340		<0.001
Oracle	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	
	Floor	Linear Regressor	XGB Ranker	Logistic Classifier	NN Ranker	XGB Regressor	NN Regressor	XGB Classifier	NN Classifier	Oracle

Figure 2. Pairwise statistical comparison of model performance using the Wilcoxon signed-rank test across 13 rolling evaluation years (2010–2022). Each cell compares two models using yearly paired cost capture at $k = 10\%$ (top 10% highest-cost patients). Shading indicates significance levels: white ($p \geq 0.05$), light gray ($0.01 \leq p < 0.05$), and black ($p < 0.01$). No multiple-testing correction is applied.

Model	Cost Capture	Δ vs NN BCE Classifier	p-value
Two-Stage XGB	0.399	-0.014	0.000
Two-Stage NN	0.407	-0.006	0.027
NN Top-K	0.391	-0.022	0.000
NN Focal	0.411	-0.002	0.414

Table 2. Mean cost capture performance across rolling-year evaluations (2010–2022).

Differences and paired Wilcoxon signed-rank tests are computed relative to the baseline neural network classifier trained with BCE loss. Shading indicates statistical significance: white $p \geq 0.05$, gray $0.01 \leq p < 0.05$, black $p < 0.01$.

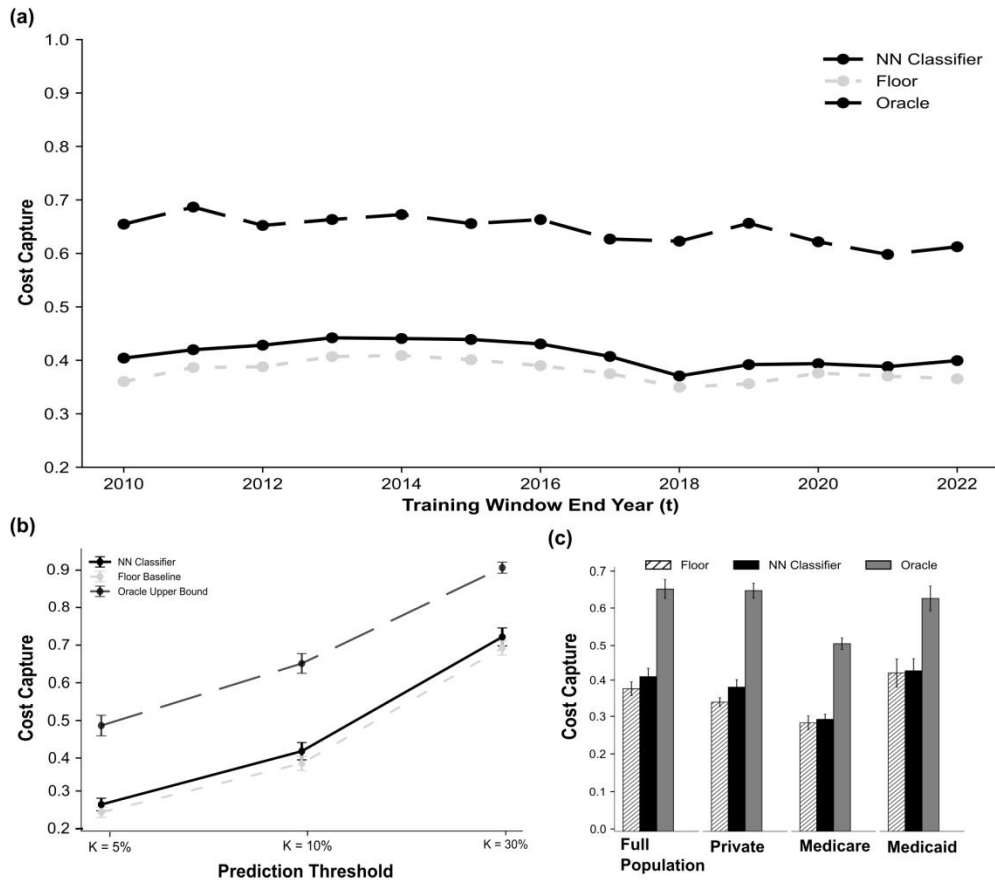


Figure 3. Stability and sensitivity of Cost Capture. Results are shown for the neural network classification model, a floor baseline computed using only prior-year expenditure, and an oracle prediction upper bound computed using realized total expenditure in the evaluation year. (a) Temporal performance stability across evaluation years. Cost Capture is computed at $k = 10\%$, corresponding to prediction of the top 10% highest-cost patients in the following year. The x-axis denotes the forecast origin (training window end year t), where models are trained using all data from 2000 through year t to predict total healthcare expenditure in year $t+1$. (b) Sensitivity of cost capture to prediction threshold. Cost Capture is evaluated at different values of k , representing the top 5%, 10%, and 30% highest-cost patients in the following year. (c) Population sensitivity of cost capture performance across MEPS sub-populations. Panels correspond to full population ($n = 257,143$), private insurance ($n = 151,248$), Medicare ($n = 56,406$), and Medicaid ($n = 43,878$) groups. Cost Capture is computed at $k = 10\%$.

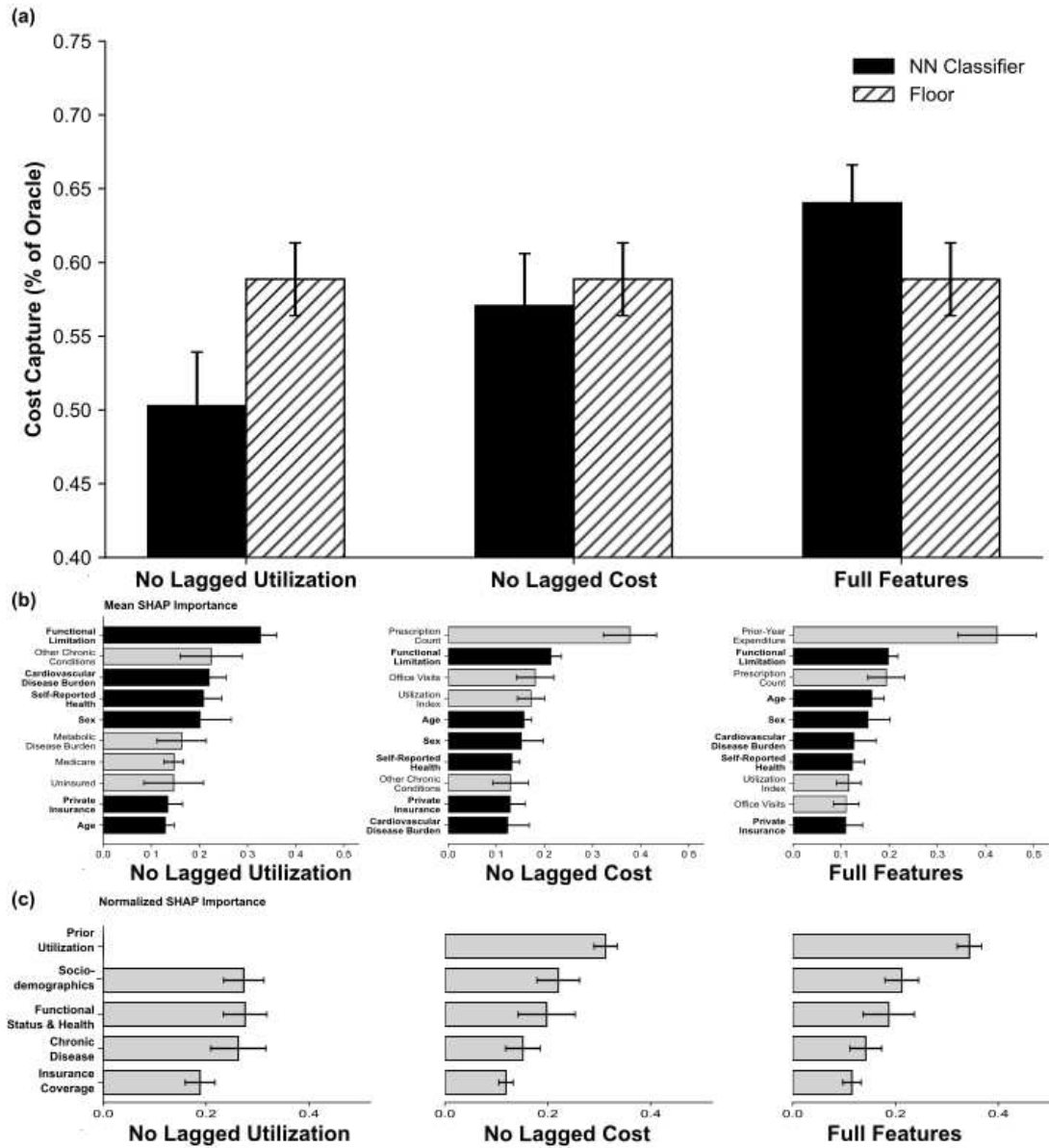


Figure 4. Feature impact on cost capture and feature importance. Three feature sets are evaluated: No Lagged Utilization (removes prior-year utilization measures including prescriptions, office visits, emergency room visits, and hospitalizations, as well as prior-year total healthcare expenditure), No Lagged Cost (removes prior-year total healthcare expenditure only), and Full Features (all features included). **(a)** Feature impact on cost capture relative to an oracle benchmark. Cost Capture is computed at $k = 10\%$, corresponding to prediction of the top 10% highest-cost patients in the following year. The floor baseline is computed using only prior-year expenditure, and the oracle prediction upper bound is computed using realized total expenditure in the evaluation year. **(b)** SHAP-based feature importance rankings across feature restriction scenarios. Panels show the top 10 predictors ranked by mean absolute SHAP value averaged across evaluation years (2010–2022), with error bars showing standard deviations. Features shown in black and bold represent those consistently ranked among the top 10 across all three scenarios. **(c)** SHAP importance aggregated by feature domain across feature restriction scenarios. Bars represent normalized mean absolute SHAP importance aggregated across feature groups and averaged across evaluation years (2010–2022), with error bars showing standard deviations. Prior Utilization includes prior-year healthcare expenditure, office visits, emergency room visits, hospitalizations, prescription utilization, and the utilization index. Sociodemographics includes age, sex, poverty category, education, region, and race/ethnicity. Functional Status & Health includes self-reported health, functional limitation indicators, mental health measures, and exercise variables. Chronic Disease includes cardiovascular, metabolic, respiratory, and other chronic disease burden indices. Insurance Coverage includes indicators for Medicare, Medicaid/CHIP, private insurance, and uninsured status.