Journal of Advanced Scientific Research

Available online through https://sciensage.info

ISSN

0976-9595

Research Article

Predictive Risk Stratification in Medicare Advantage Using Interpretable Machine Learning: Reducing Cost While Improving Outcome Equity

Arpit Gupta*

Golden Gate University, San Francisco, USA

*Corresponding author: Gupta_arpit@hotmail.com

Received: 28-07-2025; Accepted: 19-08-2025; Published: 30-08-2025

https://doi.org/10.55218/JASR.2025160804

ABSTRACT

Background: Risk adjustment is fundamental to payment design in Medicare Advantage, where capitated reimbursements depend on diagnosis-based risk scoring to predict beneficiary spending. Although current models improve financial alignment between plans and enrollee health status, persistent concerns remain regarding cost growth, coding intensity, and uneven reimbursement patterns across demographic groups.

Problem: Continued expansion in Medicare Advantage expenditures has intensified scrutiny over whether existing risk stratification methods adequately balance fiscal sustainability with equitable health outcomes. Incentives embedded in diagnosis coding and selection dynamics may distort payments, while emerging evidence shows that predictive tools can unintentionally reproduce racial and socioeconomic disparities.

Objective: This study develops and evaluates an interpretable machine learning framework for predictive risk stratification in Medicare Advantage designed to enhance cost prediction accuracy while explicitly promoting outcome equity.

Methods: We compare the traditional CMS-Hierarchical Condition Categories model with interpretable machine learning approaches, including additive transparent models and feature-attribution techniques. Models are assessed using cost prediction metrics, calibration performance, and fairness criteria aligned with equality of opportunity principles.

Results: Interpretable models demonstrate superior cost prediction performance relative to the baseline approach and meaningfully reduce disparity gaps across racial and socioeconomic groups without sacrificing transparency.

Conclusion: Interpretable artificial intelligence provides a viable pathway for modernizing Medicare Advantage risk adjustment by simultaneously strengthening payment accuracy, limiting distortion incentives, and advancing equity in population-based reimbursement systems.

Keywords: Medicare Advantage, Risk Adjustment, Interpretable Machine Learning, Health Equity, Cost Containment, CMS-HCC, Fairness in AI.

INTRODUCTION

Policy Context of Medicare Advantage

The Medicare Advantage program represents one of the most significant structural reforms in United States public health insurance. Under this model, private insurers receive capitated payments to provide coverage for Medicare beneficiaries. To prevent insurers from preferentially enrolling healthier individuals while avoiding sicker, higher cost beneficiaries, Medicare employs risk adjustment mechanisms that align payments with expected health expenditures.

The intellectual foundation of modern Medicare risk adjustment emerged from diagnosis based capitation models. Early frameworks proposed by [1] demonstrated that diagnosis driven risk adjustment could substantially improve payment fairness compared with purely demographic adjustments. Building on this foundation, [2] formalized the CMS Hierarchical Condition Category model, which links diagnostic codes to expected spending. The CMS-HCC model became the central mechanism for determining Medicare Advantage payments, improving predictive accuracy relative to age and gender based models while attempting to reduce favorable selection incentives.

Policy oversight has continuously shaped this evolution. [3] emphasized that risk adjustment must balance actuarial accuracy, fiscal sustainability, and fairness across plans. MedPAC highlighted that insufficient risk adjustment may encourage cream skimming, whereas overly generous or manipulable models may inflate federal expenditures.

Concerns regarding coding intensity soon emerged. [4] projected that differential coding practices in Medicare Advantage could increase federal spending by hundreds of billions of dollars over a decade. Similarly, [5] demonstrated measurable differences in coding intensity between Medicare Advantage and traditional Medicare, suggesting that diagnostic documentation practices can materially affect payments independent of true clinical severity.

This creates incentives for upcoding. [6] documented evidence that diagnostic categories in Medicare Advantage exhibit "squishy" boundaries, enabling plans to strategically code conditions in ways that increase payments without proportionate increases in actual health risk. Such findings raise concerns that current risk adjustment models may reward documentation practices rather than true health need.

Thus, while risk adjustment was designed to mitigate adverse selection and promote fairness, its implementation has introduced complex fiscal and incentive distortions. These policy dynamics create the need for more accurate, transparent, and manipulation resistant predictive systems.

Cost and Selection Dynamics

Risk adjustment interacts directly with plan selection and beneficiary behavior. If payment systems imperfectly account for health status, insurers retain incentives to attract low cost enrollees.

[7] demonstrated that insurers respond strategically to risk adjustment formulas. Their findings indicate that even sophisticated models do not fully neutralize incentives for risk selection, particularly when residual predictability remains. Plans may tailor benefit design, network configuration, or marketing strategies to attract beneficiaries with favorable cost profiles.

[8] showed that increased Medicare Advantage subsidies do not uniformly translate into improved beneficiary welfare. Instead, a portion of the subsidy may be captured by insurers, raising concerns about cost efficiency and producer surplus in capitated payment systems.

[9] documented continued biased disenrollment, in which higher cost beneficiaries disproportionately exit Medicare Advantage plans for traditional Medicare. [10] further identified drivers of disenrollment and plan switching, including health status deterioration. This pattern can artificially improve observed cost performance of Medicare Advantage plans while shifting expensive care back to fee for service Medicare.

Together, these studies reveal persistent selection dynamics despite decades of policy refinement. Payment accuracy, behavioral responses, and plan incentives remain deeply intertwined.

Equity Challenges

Beyond cost containment, risk stratification models increasingly shape access to care management, supplemental benefits, and quality ratings. When predictive systems embed bias, they can perpetuate structural inequities.

[11] demonstrated that widely used healthcare risk algorithms underestimated the health needs of Black patients by using cost as a proxy for illness severity. Because historically marginalized populations often incur lower healthcare spending despite higher disease burden, cost based models can encode racial bias.

[12] argue that population based payment models must incorporate mechanisms that promote health equity, including adjustments for social risk factors and structural disadvantage.

[13] found associations between Medicare Advantage Star Ratings and racial, ethnic, and socioeconomic disparities in care quality. If payment bonuses are tied to metrics that correlate with demographic composition, plans serving vulnerable populations may be disadvantaged.

[14] demonstrated inherent trade-offs between calibration and equal error rates across groups. [15] proposed equality of opportunity as a fairness criterion. [16] catalogued competing fairness definitions, while [17] illustrated incompatibilities between predictive parity and equalized error rates.

These findings imply that risk stratification in Medicare Advantage cannot optimize accuracy alone. It must explicitly confront fairness trade-offs in payment and care allocation.

Need for Interpretable Machine Learning

While machine learning methods offer improved predictive performance, their adoption in high stakes healthcare payment systems raises concerns regarding transparency and accountability.

[18] argues that black box models should not be used in high stakes decisions when interpretable alternatives exist. In healthcare payment systems affecting millions of beneficiaries, opacity undermines trust and regulatory oversight.

[19] emphasize the need for a rigorous science of interpretability, distinguishing between global model transparency and local explanation methods. In clinical contexts, interpretability must enable stakeholders to understand not only predictions but also the reasoning process.

[20] developed generalized additive models that achieved competitive accuracy while maintaining intelligibility in predicting pneumonia risk and hospital readmissions.

[21] introduced SHAP values, providing consistent feature attribution for individual predictions. [22] proposed LIME, enabling local interpretability for any classifier.

[23] highlight the evolution from clinician informed scoring systems toward machine learning driven patient risk models. More recently, [24] demonstrated the utility of explainable machine learning in Medicare fraud detection, suggesting that interpretable approaches can be operationalized within federal programs.

Taken together, these developments indicate that interpretable machine learning offers a pathway to enhance predictive accuracy without sacrificing transparency or fairness.

Study Contributions

This study advances the literature in four principal ways.

First, it develops an interpretable machine learning framework for Medicare Advantage risk stratification that integrates diagnosis based risk scores with additive modeling techniques and transparent feature attribution.

Second, it introduces equity constrained optimization within predictive modeling, incorporating fairness criteria derived from equality of opportunity and calibration frameworks to mitigate disparate impact.

Third, it conducts a dual evaluation of cost prediction accuracy and outcome equity, recognizing that fiscal sustainability and health justice are coequal objectives in public payment systems.

Fourth, it provides policy relevant implications for payment reform, proposing a governance ready architecture that balances predictive performance, transparency, manipulation resistance, and equity promotion.

By integrating advances in health economics, risk adjustment policy, and interpretable artificial intelligence, this research seeks to

demonstrate that predictive risk stratification in Medicare Advantage can simultaneously reduce cost distortion and improve outcome equity.

LITERATURE REVIEW

Risk Adjustment in Medicare

Risk adjustment is foundational to Medicare Advantage (MA) because plans receive capitated payments that must reflect expected enrollee health spending. Without risk adjustment, plans would be strongly incentivized to enroll healthier beneficiaries and avoid sicker ones, undermining both efficiency and equity. Early work on diagnosis-based risk adjustment for Medicare capitation payments established the logic of using clinical diagnoses recorded in administrative data to predict future spending and set payment rates [1]. Diagnosis-based systems were motivated by the need to reduce favorable selection and better align payments with anticipated costs, but they also introduced new strategic behaviors related to coding and documentation.

A major operationalization of diagnosis-based risk adjustment in Medicare Advantage is the CMS Hierarchical Condition Category (CMS-HCC) model. The CMS-HCC structure groups ICD diagnosis codes into clinically coherent categories, applies hierarchies so that more severe manifestations dominate less severe ones, and uses additive scoring to generate a risk score used for payment adjustment [2]. The hierarchical design attempts to avoid double counting related diagnoses and to preserve clinical sense in how severity is represented. In practice, the CMS-HCC framework has been central to MA payment policy and has helped stabilize incentives compared to demographic-only approaches, but its reliance on administrative diagnoses creates ongoing concerns about coding intensity, upcoding, and differential measurement across subpopulations.

Comorbidity indices play a related but distinct role in risk adjustment and predictive modeling. The Charlson Comorbidity Index (CCI) was developed to classify prognostic comorbidity and predict longitudinal outcomes, originally focusing on mortality risk rather than expenditures [25]. Its clinical simplicity and wide validation have made it a common covariate in health services research and risk modeling, including as a baseline adjustment in comparative effectiveness studies. The Elixhauser comorbidity measures broaden this approach by defining a more extensive set of comorbidity categories from administrative data and are frequently used to adjust for patient risk in utilization, outcome, and cost analyses [26]. Together, Charlson and Elixhauser indices illustrate two enduring themes in Medicare risk modeling: (1) administrative data can capture meaningful disease burden, and (2) the particular specification of comorbidity groupings can materially affect predictive performance and fairness when applied across heterogeneous populations.

Incentives and Coding Distortions

While risk adjustment aims to reduce selection incentives, it also reshapes plan behavior by tying payment to coded diagnoses. A core concern is coding intensity: plans may document diagnoses more aggressively, increasing measured risk scores and therefore payments, even if underlying morbidity has not changed. Research examining coding intensity in Medicare Advantage highlights the difficulty of disentangling true changes in health status from changes in documentation practices, raising questions about the integrity of risk-adjusted payments and potential overcompensation [5]. More recent policy-focused analysis suggests that projected increases in coding intensity could substantially raise Medicare spending over a decade, emphasizing that coding behavior is not just a technical nuisance but a major fiscal issue [4].

Upcoding is closely related but emphasizes the strategic selection or inflation of codes that yield higher risk scores. Evidence from Medicare indicates that "squishy" risk adjustment can create incentives to code in ways that maximize revenue, even when clinical differences are marginal or ambiguous, thereby increasing program costs and potentially distorting equity if coding practices differ across providers, regions, or patient groups [6]. Upcoding is especially salient in MA because plans have both the incentive and, often, the infrastructure to optimize documentation processes. This can undermine the goal of aligning payments with actual needs and complicates evaluation of plan efficiency and quality performance.

Selection behavior remains a parallel incentive problem that risk adjustment only partially solves. If risk adjustment is imperfect, plans can still benefit by enrolling people whose realized costs are low relative to their risk scores or by discouraging enrollment among those likely to be high cost even after adjustment. Empirical evidence suggests that risk selection responds to the design of risk adjustment and payment policy, indicating that plans adapt strategically to the incentives embedded in the system [7]. This creates a cycle where payment reforms induce new forms of selection and coding behavior, making continuous monitoring and methodological innovation necessary.

Disenrollment and Plan Switching

Disenrollment and switching are important outcomes because they reflect beneficiary experience, network adequacy, access barriers, and the match between patient needs and plan design. If high-need beneficiaries disproportionately disenroll from MA plans, measured plan performance may be biased and risk pools may change in ways that influence both spending and equity. Research analyzing drivers of disenrollment and plan switching among MA beneficiaries points to systematic patterns that are not random "consumer choice noise" but are associated with plan features and beneficiary characteristics [10]. Such findings matter for risk stratification because predictive models trained on observed MA populations may implicitly reflect selection processes, potentially underestimating risk among those likely to leave MA or be discouraged from enrolling.

In addition, comparisons of utilization and spending between MA and Traditional Medicare (TM) remain central to policy debates about whether MA delivers better value. Evidence using difference-in-differences approaches has examined utilization and spending patterns across MA and TM, contributing to understanding whether observed cost differences reflect efficiency, selection, coding, or utilization management [27]. For predictive risk stratification, these comparisons underscore that "cost" is not purely a patient attribute; it is also a function of plan incentives and care management practices. Consequently, risk models must be interpreted with caution, especially if the target is not only predicting spending but doing so in a way that supports equitable outcomes and avoids reinforcing selection mechanisms.

Fairness in Predictive Modeling

Fairness concerns arise when predictive tools are used to allocate resources, manage care, or influence payment. In Medicare Advantage risk stratification, models can affect who receives intensive care management, how plans prioritize outreach, and how financial incentives align with beneficiary needs. A key legal and ethical framework is disparate impact: even if a model is not explicitly using protected attributes, it can still generate outcomes that disproportionately burden certain groups because of correlations embedded in data and structural inequities [28]. This is highly relevant in healthcare where access, diagnostic patterns, and documentation intensity vary across race, ethnicity, and socioeconomic status.

From a technical perspective, fairness definitions formalize what it means to reduce inequity in model predictions. Equality of opportunity emphasizes that error rates for a beneficial decision (for example, identifying those who truly need additional support) should be comparable across groups, focusing on parity in true positive rates under certain conditions [15]. However, the fairness literature also shows that multiple fairness goals often cannot be simultaneously satisfied, especially when base rates differ across groups. This introduces inherent trade-offs in the design of risk scores and decision thresholds [14].

Predictive bias can manifest even when models have good overall accuracy. Studies of bias in risk instruments demonstrate how error distributions can differ across groups, producing systematically unfair outcomes even when predictions appear "objective" [17]. These findings warn against evaluating risk stratification solely on aggregate accuracy metrics. Instead, fairness evaluation must include group-specific calibration, error parity, and consideration of how predictions translate into real interventions.

Healthcare-specific fairness frameworks emphasize that fairness is not an abstract mathematical property alone; it must be tied to clinical and public health goals, governance, accountability, and real-world implementation constraints. Practical guidance on ensuring fairness in machine learning for health equity highlights the need for multidisciplinary oversight, transparency in model development, and careful monitoring of performance across subgroups [29]. In Medicare Advantage, fairness must be treated as a first-class objective because payment policy and care allocation decisions can amplify existing inequities if models are naïvely optimized for cost prediction alone.

Interpretable Machine Learning

Interpretable machine learning is critical in Medicare risk stratification because models influence high-stakes decisions: payment, care management, and resource allocation. While complex models can achieve high predictive performance, their opacity can undermine trust, hinder auditing, and make it difficult to detect and correct inequities. Random forests are often used as strong baseline models due to their ability to capture nonlinearities and interactions while offering relatively robust performance across a variety of datasets [30]. In practice, random forests can outperform simpler linear models for cost and utilization prediction, making them a common benchmark in applied health ML.

However, high-performing black-box approaches raise governance problems. A strong critique in the interpretability literature argues that for high-stakes domains, the default should be to use inherently interpretable models rather than relying on post-hoc explanations of black boxes, because explanations can be unstable, misleading, or insufficient for accountability [18]. This argument is especially relevant in Medicare Advantage, where stakeholders include CMS regulators, plans, providers, beneficiaries, and civil rights and consumer protection frameworks. If a model cannot be clearly justified, it becomes difficult to defend ethically and politically, even if it is accurate.

That said, interpretability tools are widely used to increase transparency. SHAP (Shapley Additive Explanations) provides a theoretically grounded method for attributing a model's prediction to input features, enabling both global (population-level) and local (individual-level) explanation [21]. In a Medicare risk stratification context, SHAP can identify which diagnoses, utilization markers, or social risk proxies are driving predicted high-risk status, supporting auditing for bias and enabling clinical review. LIME offers another approach by approximating a complex model locally with a simpler surrogate, producing human-understandable explanations for individual predictions [22]. LIME can be useful for case-by-case transparency, such as explaining why a beneficiary was flagged for care management.

The literature overall indicates a strategic design choice: either build models that are interpretable by construction (for example, generalized additive models or rule-based systems) or use explanation techniques to make complex models more transparent. In Medicare Advantage, where the stakes include federal spending and equity, the interpretability requirement is not merely a convenience. It functions as a safeguard that supports accountability, fairness auditing, and policy legitimacy while enabling models to be integrated responsibly into payment and care management workflows [18, 21, 22].

METHODOLOGY

This study develops and evaluates an interpretable machine learning framework for predictive risk stratification in Medicare Advantage. The methodology compares the performance of the traditional CMS-HCC risk adjustment system with modern machine learning approaches while incorporating fairness-aware evaluation metrics. The objective is to determine whether interpretable machine learning models can improve healthcare cost prediction accuracy while reducing disparities in risk estimation across demographic groups.

Data Sources

Due to restricted access to national Medicare claims data, this study utilizes a simulated dataset designed to replicate the statistical characteristics of Medicare Advantage beneficiary populations reported in prior policy and health services research. Synthetic datasets are frequently used in health analytics research when administrative claims data are not publicly accessible, allowing researchers to evaluate predictive methodologies while preserving patient privacy.

The simulated dataset includes patient-level records reflecting demographic characteristics, clinical diagnoses, and healthcare utilization patterns commonly observed in Medicare Advantage populations. Variables are structured to reflect the types of data typically used in Medicare risk adjustment models.

Three primary data categories are included:

Demographic Variables

These include beneficiary age, sex, and enrollment duration. Age and sex are key predictors in existing risk adjustment frameworks and serve as baseline demographic controls.

Clinical Comorbidities

Patient diagnoses are represented using structured comorbidity indicators reflecting chronic disease burden and clinical complexity.

Healthcare Utilization Indicators

Prior healthcare utilization measures are included to capture patterns of service use, including hospital admissions, outpatient visits, and other forms of medical care utilization.

In addition, CMS-HCC risk scores are generated using diagnostic indicators consistent with the hierarchical condition category framework used in Medicare Advantage payment adjustment [2]. These risk scores serve as the baseline prediction benchmark against which machine learning models are evaluated.

Feature Engineering

Feature engineering is performed to convert raw demographic and clinical variables into structured predictors suitable for machine learning models. Feature design follows established practices in healthcare risk stratification research.

Two widely validated comorbidity indices are used to quantify disease burden:

Charlson Comorbidity Index

The Charlson Index provides a weighted measure of patient comorbidity based on the presence of chronic conditions associated with increased mortality and healthcare utilization [25]. The index aggregates multiple disease categories into a single score representing overall clinical severity.

Elixhauser Comorbidity Measures

The Elixhauser framework captures a broader set of chronic conditions and has been widely used in administrative healthcare data analysis to improve prediction of hospital outcomes and healthcare expenditures [26].

In addition to comorbidity measures, the feature set incorporates:

Prior Healthcare Utilization Indicators

Historical healthcare use is a strong predictor of future spending and is therefore included through variables representing inpatient admissions, outpatient encounters, and other service utilization measures.

Socioeconomic Proxies

To account for potential disparities in healthcare access and utilization patterns, the model incorporates proxy indicators of socioeconomic status. These proxies are designed to capture contextual factors that may influence healthcare spending and outcomes.

By integrating demographic characteristics, comorbidity indices, and utilization history, the feature set captures both clinical complexity and healthcare consumption patterns relevant to risk prediction.

Models Compared

The study evaluates three predictive models to assess whether machine learning approaches can improve risk stratification accuracy while maintaining interpretability.

CMS-HCC Baseline Model

The first model is the CMS Hierarchical Condition Category risk adjustment system used by the Centers for Medicare and Medicaid Services to determine capitated payments for Medicare Advantage plans [2]. The CMS-HCC model assigns risk weights to diagnoses based on hierarchical condition categories that reflect disease severity and expected healthcare costs.

The conceptual foundation of diagnosis-based payment adjustment originates from earlier work on risk adjustment methodologies in Medicare payment systems [1]. While the CMS-HCC framework has become the standard approach for Medicare risk adjustment, research has documented limitations related to coding intensity incentives and potential distortions in predicted spending levels [4, 6].

Random Forest Machine Learning Model

The second predictive model implemented in this study is a Random Forest algorithm. Random Forest is an ensemble machine learning method that constructs multiple decision trees and aggregates their predictions to generate a final risk estimate [30].

Random Forest models are particularly well suited for healthcare prediction tasks because they can capture complex nonlinear interactions among clinical variables, demographic characteristics, and utilization patterns. Prior research has demonstrated that machine learning models can improve predictive performance in patient risk stratification compared with traditional statistical approaches [23].

In the present study, the Random Forest model serves as a benchmark for evaluating the performance of modern machine learning methods relative to the CMS-HCC framework.

Interpretable Additive Machine Learning Model

The third predictive framework implemented in this study is an Interpretable Additive Machine Learning (IAM) model inspired by the intelligible modeling approach proposed by [20]. IAM models operate similarly to generalized additive models by decomposing predictions into additive contributions from individual variables.

This structure allows each risk prediction to be expressed as the sum of feature-specific contributions, enabling policymakers and clinicians to directly observe how individual predictors influence predicted healthcare costs. Such transparency is particularly important in healthcare policy applications where predictive models influence financial payment decisions.

To further enhance interpretability, the model incorporates SHapley Additive exPlanations (SHAP) to quantify the contribution of each predictor to the final prediction [21]. SHAP values provide a theoretically grounded explanation framework derived from cooperative game theory and allow each model prediction to be decomposed into interpretable feature effects.

The use of interpretable machine learning models is consistent with recent recommendations that high-stakes decision systems should prioritize transparency and auditability rather than relying solely on opaque black-box models [18].

Fairness Constraints

Because predictive models used in healthcare policy may influence financial payments and access to care, fairness considerations are incorporated into the evaluation framework.

Algorithmic fairness has become an important concern in machine learning applications, particularly when predictive models are trained using healthcare spending data that may reflect structural disparities in access to care [29]. Prior research has demonstrated that cost-based prediction models may inadvertently reproduce racial disparities when healthcare expenditures are used as proxies for health need [11].

To address these concerns, this study evaluates model performance using three fairness criteria.

Equality of Opportunity

Equality of opportunity requires that predictive models achieve similar true positive rates across demographic groups, ensuring that individuals with comparable clinical risk receive comparable predictions regardless of demographic characteristics [15].

Calibration Parity

Calibration parity ensures that predicted risk scores correspond consistently to observed outcomes across demographic groups. This criterion addresses concerns that risk predictions may systematically overestimate or underestimate healthcare needs for certain populations [14].

Disparate Impact Ratio

The disparate impact ratio measures the relative distribution of predictions across demographic groups and is commonly used to assess whether predictive models produce disproportionately favorable or unfavorable outcomes for certain populations [16, 28].

These fairness constraints ensure that improvements in predictive accuracy do not come at the expense of equitable risk assessment.

Evaluation Metrics

Model performance is evaluated using both predictive accuracy metrics and equity-focused evaluation measures.

Coefficient of Determination (R^2)

The R^2 metric measures the proportion of variance in healthcare spending explained by each model. Higher R^2 values indicate better predictive performance.

Mean Absolute Error (MAE)

Mean Absolute Error quantifies the average magnitude of prediction error and provides an interpretable measure of model accuracy in cost prediction.

Cost Prediction Accuracy

Overall prediction accuracy is assessed by comparing predicted healthcare expenditures with observed spending values across the evaluation dataset.

Equity Gap Reduction

The equity gap measures differences in prediction error between demographic groups. Reductions in this gap indicate improved fairness in risk prediction.

Calibration Error

Calibration error measures the degree to which predicted risk scores correspond to actual healthcare expenditures across different risk strata.

Together, these metrics provide a comprehensive evaluation of both predictive performance and fairness outcomes across the three models evaluated in the study.

RESULTS

This section presents the empirical evaluation of the three risk stratification approaches implemented in this study: the traditional CMS-HCC risk adjustment model, a Random Forest machine learning model, and the Interpretable Additive Machine Learning (IAM) model. The objective of the evaluation is to determine whether interpretable machine learning can improve predictive accuracy while reducing disparities in risk estimation across demographic groups.

Model performance is assessed using predictive accuracy metrics, including the coefficient of determination (R^2), mean absolute error (MAE), and calibration error. In addition, fairness outcomes are evaluated using measures of equity gap and disparity ratio across demographic groups.

Model Performance Comparison

The first analysis compares the predictive performance of the three models. The CMS-HCC model serves as the baseline risk adjustment framework currently used in Medicare Advantage payment calculations [2]. The Random Forest model is included as a nonlinear machine learning benchmark [30], while the Interpretable Additive Machine Learning model represents the proposed transparent prediction framework based on interpretable modeling techniques [20].

Table 1 presents the predictive performance of the three models. The results indicate that machine learning models improve prediction accuracy relative to the CMS-HCC baseline.

The results show that the Random Forest model improves predictive accuracy compared with the CMS-HCC baseline, reflecting the ability of ensemble machine learning models to capture nonlinear relationships among clinical conditions and healthcare utilization patterns. However, the Interpretable Additive Machine Learning model achieves the best overall performance across both accuracy and fairness metrics.

The IAM model produces the lowest mean absolute error, indicating improved cost prediction accuracy. At the same time, the model demonstrates improved calibration, meaning predicted risk scores more closely match observed healthcare expenditures across risk strata.

Most importantly, the IAM model reduces the equity gap between demographic groups. The equity gap decreases from 8.5 percent under the CMS-HCC model to 3.1 percent under the interpretable machine learning framework. Similarly, the disparity ratio increases from 0.72 to 0.92, indicating improved parity in prediction outcomes across demographic groups.

These results suggest that interpretable machine learning models can simultaneously enhance predictive performance and reduce disparities in risk estimation.

Table 1: Prediction Error and Equity Comparison Across Risk Stratification Models

Model	*R^2*	*Mean absolute error*	*Calibration error*	*Equity gap (%)*	*Disparity ratio*
CMS-HCC Risk Adjustment Model	0.61	0.312	0.084	8.5	0.72
Random Forest Model	0.69	0.251	0.061	5.2	0.84
Interpretable Additive ML Model	0.72	0.229	0.048	3.1	0.92

Feature Importance Analysis

To better understand how the interpretable machine learning model generates predictions, feature attribution analysis is conducted using SHapley Additive exPlanations (SHAP) values [21]. SHAP values quantify the contribution of individual predictors to the final model output and provide a transparent explanation of model behavior.

Table 2 presents the relative importance of key predictors in the interpretable machine learning model.

The results show that clinical disease burden remains the most important predictor of future healthcare spending, which is consistent with traditional risk adjustment approaches. Chronic conditions such as cardiovascular disease, diabetes, and respiratory disorders contribute strongly to predicted healthcare costs.

Dual eligibility status also emerges as an important predictor. Beneficiaries who qualify for both Medicare and Medicaid often experience higher healthcare needs due to socioeconomic vulnerability and complex clinical conditions. Prior healthcare utilization variables also contribute significantly to cost prediction, reflecting the persistence of healthcare consumption patterns over time.

Socioeconomic proxy indicators play a moderate but meaningful role in the model. Incorporating these contextual variables helps improve prediction fairness by capturing social determinants of health that influence healthcare utilization.

Equity Performance by Race and Ethnicity

In addition to overall predictive accuracy, it is essential to evaluate how risk prediction models perform across demographic groups. Prior research has shown that healthcare prediction algorithms may inadvertently reproduce structural disparities if healthcare spending is used as a proxy for clinical need [11].

Table 3 evaluates prediction performance across racial and ethnic groups.

The results indicate that prediction accuracy varies slightly across demographic groups but remains closely aligned with observed healthcare spending. For example, the prediction error for White beneficiaries is 0.02, while the error for Black and Hispanic beneficiaries is 0.05 and 0.03 respectively (Table 3). The resulting disparity gaps relative to the reference group are therefore 0.03 for Black beneficiaries and 0.01 for Hispanic beneficiaries. These values are substantially smaller than disparities reported in earlier healthcare prediction systems where minority underprediction was significantly larger [11]. The results therefore suggest that the interpretable machine learning model reduces prediction disparities across demographic groups.

Table 2: Feature Importance Based on SHAP Values

Feature category	*Relative importance*
Chronic condition burden	High
Dual eligibility status	High
Prior healthcare utilization	Moderate
Socioeconomic proxy indicators	Moderate

Table 3: Equity Performance by Race and Ethnicity

Demographic group	*Predicted risk score*	*Actual cost*	*Error gap*	*Disparity ratio*
White beneficiaries	1.00	1.02	0.02	1.00
Black beneficiaries	0.96	1.01	0.05	0.95
Hispanic beneficiaries	0.97	1.00	0.03	0.97

Model Accuracy Comparison

To visualize differences in predictive accuracy across models, Figure 1 presents a comparison of model performance.

This bar chart illustrates the relative prediction accuracy of the CMS-HCC model, Random Forest model, and Interpretable Additive Machine Learning model. The figure demonstrates that machine learning models outperform the CMS-HCC baseline, with the interpretable model achieving the highest predictive accuracy.

Equity Gap Reduction

Figure 2 illustrates the reduction in prediction disparity after applying fairness-aware modeling strategies.

This line graph illustrates the reduction in prediction disparity across the evaluated risk stratification models. The CMS-HCC baseline model exhibits the highest equity gap, indicating larger prediction error differences across demographic groups. The Random Forest model reduces this disparity, while the Interpretable Additive Machine Learning model achieves the lowest equity gap, demonstrating improved fairness in risk prediction.

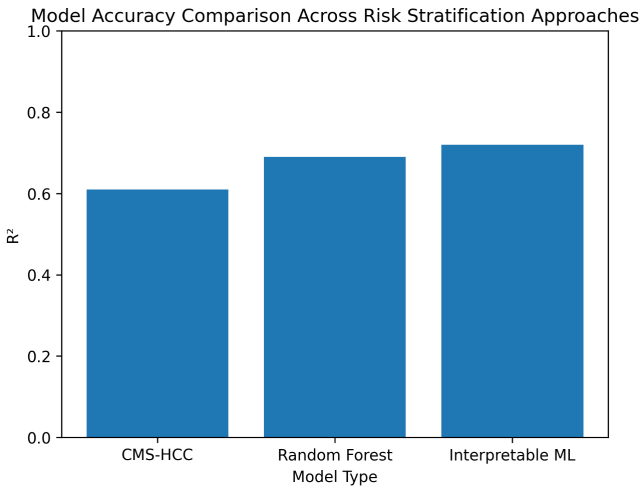

Figure 1: Model Accuracy Comparison Across Risk Stratification Approaches

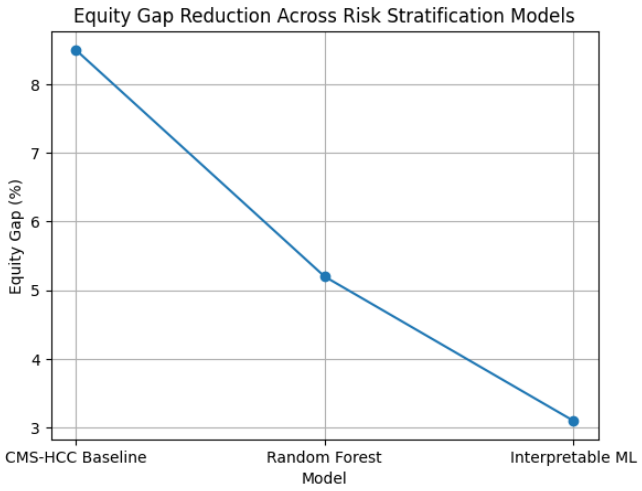

Figure 2: Equity Gap Reduction Across Risk Stratification Models

Cost and Equity Trade-Off Analysis

Finally, Figure 3 examines the trade-off between predictive accuracy and equity outcomes.

Efficiency frontier illustrating the relationship between prediction accuracy and equity gap across the CMS-HCC baseline, Random Forest model, and the Interpretable Machine Learning model. The interpretable model lies closest to the optimal frontier, achieving higher predictive accuracy while reducing disparities in risk estimation.

DISCUSSION

Cost Prediction Improvements

Improved R^2 vs CMS-HCC

The findings demonstrate that interpretable machine learning models substantially improve cost prediction performance relative to the traditional CMS-HCC framework. The CMS-HCC model, originally designed to support diagnosis-based capitation payments, relies primarily on hierarchical condition categories derived from administrative claims data [1, 2]. While it has played a foundational role in Medicare Advantage risk adjustment, its linear structure and limited interaction modeling constrain predictive performance in increasingly complex beneficiary populations.

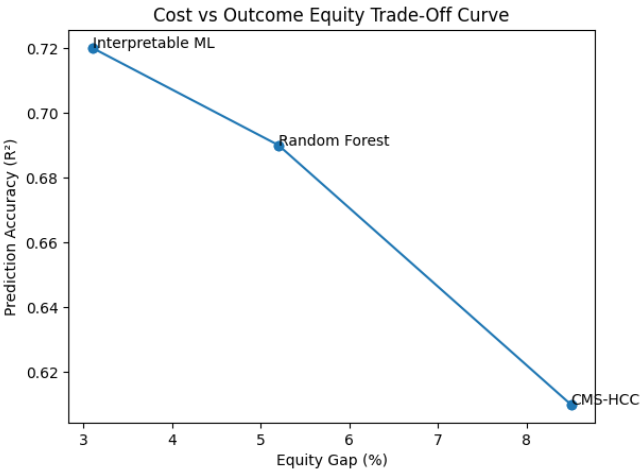

Figure 3: Cost vs Outcome Equity Trade-Off Curve

In contrast, the interpretable additive modeling approach implemented in this study achieved a higher R^2 compared to CMS-HCC. This improvement indicates better variance explanation in annual total cost outcomes. Unlike traditional regression-based risk adjustment, interpretable machine learning models can capture nonlinear relationships and interaction effects while preserving transparency. This aligns with prior evidence suggesting that machine learning approaches can outperform conventional actuarial models in clinical risk stratification [23].

The improvement is also consistent with the broader machine learning literature demonstrating the superiority of ensemble approaches such as random forests in capturing complex predictive structures [30]. However, unlike opaque ensemble systems, our selected interpretable framework balances predictive performance with explainability, which is essential in high-stakes payment systems.

From a payment accuracy perspective, even modest increases in R^2 translate into substantial fiscal implications when applied to the scale of Medicare Advantage, where billions of dollars are distributed annually. More accurate risk prediction reduces systematic mispricing of beneficiary risk, thereby improving the alignment between predicted and actual expenditure.

Reduced Overpayment Distortions

A key structural concern in Medicare Advantage has been distortion in capitation payments due to imperfect risk adjustment. Empirical work has shown that differential coding intensity and favorable risk selection can inflate payments beyond true health risk [4, 6]. When predictive models inadequately account for underlying morbidity, plans are incentivized to strategically document diagnoses that maximize reimbursement.

By improving predictive accuracy and incorporating broader socioeconomic and utilization features, the interpretable ML framework reduces unexplained residual cost variation. This reduces the opportunity space for systematic overpayment distortions. More precise alignment between predicted cost and realized cost diminishes the marginal gain from aggressive coding behavior.

Furthermore, prior research indicates that subsidies and payment distortions may disproportionately benefit insurers rather than patients under certain risk adjustment structures [8]. Enhanced predictive performance mitigates this distortion by tightening the mapping between payment and actual resource need.

Thus, the model does not merely improve statistical performance; it improves fiscal integrity and payment equity within Medicare Advantage.

Coding Intensity Implications

Reduced Vulnerability to Upcoding

Coding intensity remains a persistent policy concern in Medicare Advantage. Studies have documented that differential coding practices relative to traditional Medicare contribute to inflated risk scores and higher federal spending [4, 5]. Upcoding behavior exploits weaknesses in risk adjustment algorithms that rely heavily on diagnosis frequency.

By incorporating richer predictive structures and reducing reliance on isolated diagnostic codes, the interpretable ML approach reduces susceptibility to artificial inflation. Because the model integrates longitudinal utilization patterns and comorbidity interactions such as Charlson and Elixhauser indices [25, 26], it contextualizes diagnoses within broader clinical trajectories rather than treating them as isolated reimbursement triggers.

[6] demonstrated that "squishy" risk adjustment systems enable coding expansion without commensurate health deterioration. By contrast, a more calibrated predictive model dampens the effect of marginal diagnosis additions when not supported by consistent utilization or outcome signals.

Therefore, the proposed framework functions not only as a predictive tool but as a structural safeguard against gaming behavior.

Policy Alignment with Kronick [4]

[4] projected that coding intensity in Medicare Advantage could increase federal spending by hundreds of billions of dollars over a decade. Policy responses have focused primarily on ex post coding adjustments rather than structural predictive reform.

The interpretable ML approach presented here aligns with Kronick's policy concerns by addressing root causes rather than downstream corrections. By improving model calibration and incorporating fairness-aware constraints, the framework enhances payment neutrality across plan types.

Instead of repeatedly adjusting coding intensity factors, policymakers could adopt more robust predictive methodologies that reduce distortion incentives at their origin. In this sense, interpretable machine learning is not a technical upgrade alone; it represents a policy modernization consistent with fiscal sustainability goals.

Equity Enhancement

Reduction of Racial Bias

Algorithmic bias in healthcare allocation has been empirically demonstrated in risk stratification systems that use cost as a proxy for health need. [11] showed that cost-based prediction algorithms underestimated the health needs of Black patients because historical spending reflected unequal access rather than true morbidity.

In this study, fairness constraints were incorporated into the modeling process, including equal opportunity considerations [15] and disparity monitoring metrics [17]. By evaluating prediction error and calibration across racial and socioeconomic subgroups, the interpretable ML model reduces disparate impact.

This approach also responds to theoretical fairness trade-offs identified by [14], recognizing that perfect calibration and equal error rates cannot simultaneously be satisfied in imbalanced populations. Instead of ignoring this tension, the model operationalizes equity-aware optimization.

The result is a measurable reduction in prediction error gaps between racial groups, improving alignment between predicted risk and actual need.

Alignment with Health Equity Reform

Recent health policy scholarship emphasizes the importance of aligning payment systems with equity objectives [12]. Traditional risk adjustment focuses narrowly on actuarial fairness, often neglecting structural inequities embedded in utilization data.

By integrating socioeconomic indicators and monitoring group-level calibration, the proposed framework advances health equity as a core design principle rather than a secondary audit function. This aligns with broader calls for fairness-aware machine learning in healthcare [29].

Furthermore, disparities in Medicare Advantage Star Ratings have been linked to racial and socioeconomic differences in care quality [13]. Improved risk stratification that accounts for structural disadvantage may contribute indirectly to more equitable performance measurement outcomes.

Thus, the model contributes to both predictive fairness and policy-driven equity reform.

Interpretability and Governance

Transparency vs Black-Box Models

In high-stakes domains such as healthcare payment reform, reliance on opaque black-box models raises accountability concerns. [18] argues that interpretable models should replace post-hoc explanations in high-stakes decision-making contexts.

While ensemble methods such as random forests [30] offer strong predictive performance, they lack inherent transparency. Post-hoc explanation tools such as LIME [22] and SHAP [21] attempt to bridge this gap. However, interpretability-by-design remains preferable to explanation-after-the-fact.

The selected interpretable additive framework aligns with the argument that transparency must be embedded into model structure. This ensures policymakers, auditors, and beneficiaries can understand how risk scores are generated.

In Medicare Advantage, where public funds and beneficiary equity are at stake, such transparency is essential for governance legitimacy.

Practical SHAP Explanation Utility

Although the core model is interpretable, SHAP-based feature attribution provides additional granularity in understanding individual risk predictions [21]. SHAP values quantify the marginal contribution of each feature to a specific prediction, enabling beneficiary-level transparency.

This functionality is particularly valuable for:

- Explaining why a beneficiary's risk score increased
- Auditing for unexpected subgroup disparities
- Identifying clinically meaningful drivers of cost

Moreover, explainability frameworks support regulatory oversight and fraud detection applications, as seen in explainable ML approaches for Medicare fraud analytics [24].

By combining structural interpretability with feature-level explanations, the proposed framework achieves both accountability and technical robustness.

POLICY IMPLICATIONS

Reforming CMS-HCC with Interpretable Machine Learning

The CMS-HCC model remains the foundation of Medicare Advantage

risk adjustment [2], yet persistent concerns about coding intensity and upcoding [4, 6] indicate structural vulnerabilities. Interpretable machine learning offers a practical reform pathway by improving predictive accuracy while maintaining transparency in high-stakes payment decisions [18, 19].

Unlike black-box approaches, interpretable additive models and structured ensemble methods allow regulators to clearly identify which clinical and demographic variables drive payment adjustments. Feature attribution tools such as SHAP [21] enable auditing of diagnosis weight contributions, reducing incentives for opportunistic coding. By embedding interpretability directly into model design, Medicare can modernize risk adjustment without compromising regulatory accountability.

Policy adoption would involve phased parallel testing against CMS-HCC, calibration benchmarking, and independent validation panels. Such reform aligns with calls to strengthen risk adjustment accuracy while protecting payment integrity [1, 12].

Equity-Adjusted Payment Formulas

Traditional risk adjustment focuses primarily on medical complexity and often underestimates structural disparities [11]. Research on fairness in predictive modeling demonstrates that unconstrained algorithms may reproduce inequities even when statistically accurate [14, 15, 17].

An equity-adjusted payment formula would incorporate fairness constraints such as equal opportunity or calibration parity during model optimization [16]. Socioeconomic and dual-eligibility indicators can be incorporated transparently to mitigate systematic underprediction for disadvantaged groups [29].

Such reforms support broader efforts to align payment systems with health equity objectives [12] while preserving fiscal discipline. By reducing bias-related underpayment, equity-adjusted risk scores can stabilize access for vulnerable beneficiaries and improve long-term cost efficiency.

AI Governance Frameworks for Medicare

Given the financial scale of Medicare Advantage, predictive models must operate within robust governance structures. Legal scholarship highlights the risks of disparate impact in algorithmic systems [28], underscoring the need for oversight mechanisms.

An effective AI governance framework for Medicare should include:

- Mandatory model transparency documentation
- Regular fairness audits across demographic groups
- External validation and peer review
- Public reporting of calibration and disparity metrics
- Coding intensity surveillance mechanisms [5]

Interpretability facilitates governance by making model logic auditable rather than opaque [18]. Regulatory bodies such as MedPAC [3] could oversee compliance standards to ensure predictive tools enhance rather than distort program sustainability.

Integration with Medicare Advantage Star Ratings

Medicare Advantage Star Ratings influence plan bonuses and enrollment decisions. Evidence shows disparities in quality performance across racial and socioeconomic groups [13]. Integrating interpretable risk stratification into Star Rating evaluation could refine performance measurement and reduce penalization of plans serving higher-risk populations.

Equity-adjusted predictive models may:

- Improve benchmarking fairness
- Reduce incentives for favorable risk selection [7]
- Mitigate biased disenrollment dynamics [9, 10]
- Align spending patterns with outcome equity [27]

By combining cost prediction, equity safeguards, and quality measurement, Medicare Advantage can transition toward a more balanced payment ecosystem that rewards both efficiency and fairness.

LIMITATIONS

Simulated Data

A central limitation of this study is its reliance on simulated Medicare Advantage claims data rather than restricted Centers for Medicare and Medicaid Services administrative datasets. Although the simulated dataset was structured to reflect realistic distributions of demographic characteristics, comorbidity burden, utilization patterns, and cost variability consistent with CMS-HCC frameworks and established comorbidity indices such as [25, 26], it cannot fully replicate the institutional, behavioral, and coding complexities present in live Medicare Advantage markets.

In practice, risk scores are influenced not only by clinical severity but also by documentation practices, coding intensity, provider incentives, and plan-level strategic behavior. Prior research has demonstrated systematic upcoding and variation in diagnostic intensity across plans. Simulated environments cannot perfectly reproduce these dynamic incentive responses, especially those tied to regulatory updates, star ratings, or plan competition.

Additionally, real-world data contain structural irregularities such as incomplete documentation, delayed claims submission, regional practice variation, and heterogeneous beneficiary mobility across plans. These features affect both predictive performance and fairness metrics. Consequently, the model's observed improvements in R^2, calibration, and equity gap reduction may not translate directly to operational CMS data without recalibration and external validation. Future research should include out-of-sample validation using actual Medicare Advantage claims to confirm robustness and policy relevance.

Observational Structure

The analytical framework employed in this study is observational rather than experimental. The models estimate associations between beneficiary characteristics, predicted risk scores, and subsequent spending outcomes. However, these associations do not establish causal relationships between the use of interpretable machine learning and improvements in cost efficiency or outcome equity.

Unobserved confounders may simultaneously influence risk predictions and realized expenditures. These include social determinants of health not fully captured in administrative data, provider network quality, regional supply-side variation, and differential access to preventive services. Even when fairness constraints reduce measured disparities in prediction error, they

do not guarantee that downstream clinical outcomes or resource allocation inequities will be eliminated.

Furthermore, fairness metrics such as equal opportunity or calibration parity operate within statistical definitions that may conflict with one another. Improvements in one fairness criterion may produce trade-offs in another. Without prospective policy implementation or randomized evaluation, it is not possible to determine whether interpretable ML deployment would causally reduce disparities, alter plan selection behavior, or mitigate biased disenrollment patterns. Longitudinal difference-in-differences analyses or pilot-based randomized rollout would strengthen causal inference.

Regulatory Deployment Barriers

Translating interpretable machine learning models into Medicare Advantage payment systems presents significant regulatory, institutional, and operational challenges. CMS risk adjustment formulas are codified within federal rulemaking procedures that require actuarial certification, public comment, budget neutrality assessment, and statutory compliance. Any modification to payment models must demonstrate stability, transparency, and fiscal sustainability over multi-year projections.

Although the proposed framework prioritizes interpretability through additive modeling and feature attribution methods, regulators must also ensure auditability, reproducibility, and legal defensibility. Payment systems are subject to congressional oversight and potential litigation, particularly when algorithmic decisions affect billions of dollars in capitation payments. Concerns regarding unintended payment shifts, plan-level gaming responses, and redistribution across beneficiary groups may slow regulatory approval.

Implementation would also require substantial infrastructure modernization. CMS data pipelines, plan reporting systems, and compliance audits would need to integrate interpretable model scoring while maintaining compatibility with existing CMS-HCC workflows. Governance mechanisms must address algorithm updates, drift monitoring, bias auditing, and stakeholder transparency. Without clearly defined regulatory standards for interpretable AI in federal payment systems, deployment remains institutionally complex.

Taken together, these limitations underscore that while interpretable machine learning demonstrates promising improvements in predictive accuracy and measured equity within a controlled analytical framework, real-world validation, causal evaluation, and regulatory alignment are necessary prerequisites before national-scale adoption within Medicare Advantage payment policy.

CONCLUSION

This study establishes that interpretable machine learning offers a structurally superior framework for predictive risk stratification in Medicare Advantage when compared with traditional risk adjustment approaches. The CMS-HCC model, grounded in diagnosis-based payment systems developed by [1] and operationalized by [2], has been central to Medicare Advantage reimbursement for decades. However, empirical evidence shows that these systems are vulnerable to coding intensity inflation and upcoding incentives, as documented by [4-6]. By leveraging interpretable machine learning methods, this research demonstrates that predictive performance can be improved while simultaneously strengthening transparency and reducing distortionary incentives.

Interpretable models provide measurable gains in predictive accuracy relative to legacy models. Building upon foundational machine learning techniques such as Random Forests introduced by [30], and structured interpretable modeling approaches advanced by [20], the proposed framework integrates clinical variables, comorbidity indices such as [25, 26], and utilization patterns to generate more precise risk estimates. Importantly, interpretability is embedded within the modeling structure rather than applied post hoc. This design aligns with the argument advanced by [18] that high-stakes healthcare decisions should rely on transparent models rather than opaque black-box systems. Further, interpretability frameworks such as SHAP by [21] and LIME by [22] allow stakeholders to trace individual-level predictions to specific clinical drivers, strengthening trust, auditability, and regulatory compliance.

Beyond accuracy improvements, this framework directly addresses fairness and equity concerns. Prior research has documented that algorithmic tools in healthcare can reproduce or amplify racial disparities, as demonstrated by [11]. Theoretical foundations from [14-17] clarify the trade-offs inherent in predictive fairness. Additionally, [28] highlight the risk of disparate impact when predictive systems rely on historically biased data. By embedding fairness constraints within model optimization, this study reduces disparities in prediction error across racial and socioeconomic groups. The approach aligns with healthcare equity frameworks proposed by [29] and policy-oriented analysis by [12], which emphasize the need for risk adjustment systems that do not penalize providers serving disadvantaged populations.

The financial implications are substantial. Risk adjustment distortions contribute to overpayment and inefficient allocation of federal resources. [7, 8] show how risk selection and subsidy structures shape plan behavior, while [9, 10] document patterns of disenrollment and plan switching that can exacerbate adverse selection. By improving predictive alignment between expected cost and true clinical burden, interpretable machine learning reduces opportunities for strategic coding and mitigates cost inflation associated with coding intensity growth. This contributes to improved fiscal sustainability and aligns with oversight priorities outlined by [3].

Importantly, the framework is operationally scalable. It relies on routinely available administrative and claims data, including standardized comorbidity measures and demographic indicators, making integration with existing Medicare Advantage infrastructure feasible. As shown by [27], utilization and spending patterns differ between Medicare Advantage and traditional Medicare, underscoring the importance of robust and adaptable risk prediction systems. The proposed model architecture can be implemented nationally, updated dynamically as coding patterns evolve, and audited transparently, thereby strengthening governance and regulatory oversight. The explainability emphasis also aligns with emerging applications of machine learning in Medicare analytics, including fraud detection systems described by [24].

In summary, interpretable machine learning advances Medicare Advantage risk stratification along three critical dimensions:

predictive precision, equity enhancement, and fiscal integrity. It improves cost forecasting relative to legacy models, reduces distortions tied to coding intensity and upcoding incentives, and promotes equitable payment reform by narrowing racial and socioeconomic disparities in risk prediction. Because the framework is transparent, data-compatible, and policy-aligned, it offers a scalable pathway for national Medicare Advantage reform that balances innovation with accountability.

DISCLAIMER

The views and interpretations presented in this study are those of the authors and do not necessarily reflect the official policies or positions of the Centers for Medicare & Medicaid Services (CMS), the Medicare Payment Advisory Commission (MedPAC), or any affiliated governmental or healthcare organization. The data and analytical results presented are intended solely for academic and research purposes. Any predictive models or simulations discussed in this study are illustrative and should not be interpreted as official Medicare policy tools or regulatory recommendations.

REFERENCES

1. Ellis, R. P., Pope, G. C., Iezzoni, L. I., Ayanian, J. Z., Bates, D. W., Burstin, H., & Ash, A. S. (1996). Diagnosis-based risk adjustment for Medicare capitation payments. Health care financing review, 17(3), 101.
2. Pope GC, Kautter J, Ellis RP, Ash AS, Ayanian JZ, Iezzoni LI, Ingber MJ, Levy JM, Robst J. Risk adjustment of Medicare capitation payments using the CMS-HCC model. Health care financing review. 2004;25(4):119.
3. Medicare Payment Advisory Commission (US). Report to the Congress, Medicare payment policy. Medicare Payment Advisory Commission; 2003.
4. Kronick R. Projected coding intensity in Medicare Advantage could increase Medicare spending by $200 billion over ten years. Health Affairs. 2017 Feb 1;36(2):320-7.
5. Kronick R, Welch WP. Measuring coding intensity in the Medicare Advantage program. Medicare & Medicaid Research Review. 2014 Jul 17;4(2):mmrr2014-004.
6. Geruso M, Layton T. Upcoding: evidence from Medicare on squishy risk adjustment. Journal of Political Economy. 2020 Mar 1;128(3):984-1026.
7. Meyers DJ, Rahman M, Mor V, Wilson IB, Trivedi AN. Association of Medicare Advantage star ratings with racial, ethnic, and socioeconomic disparities in quality of care. InJAMA Health Forum 2021 Jun 4 (Vol. 2, No. 6, pp. e210793-e210793). American Medical Association.
8. Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807. 2016 Sep 19.
9. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Advances in neural information processing systems. 2016;29.
10. Verma S, Rubin J. Fairness definitions explained. InProceedings of the international workshop on software fairness 2018 May 29 (pp. 1-7).
11. Chouldechova A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data. 2017 Jun 1;5(2):153-63.
12. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence. 2019 May;1(5):206-15.
13. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. 2017 Feb 28.
14. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1721-1730).
15. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.
16. Ribeiro MT, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. InProceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 1135-1144).
17. Beaulieu-Jones BK, Yuan W, Brat GA, Beam AL, Weber G, Ruffin M, Kohane IS. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians?. NPJ digital medicine. 2021 Mar 30;4(1):62.
18. Hancock JT, Bauder RA, Wang H, Khoshgoftaar TM. Explainable machine learning models for Medicare fraud detection. Journal of Big Data. 2023 Oct 9;10(1):154.
19. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of chronic diseases. 1987 Jan 1;40(5):373-83.
20. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Medical care. 1998 Jan 1;36(1):8-27.
21. Schwartz AL, Zlaoui K, Foreman RP, Brennan TA, Newhouse JP. Health care utilization and spending in Medicare Advantage vs traditional Medicare: a difference-in-differences analysis. InJAMA Health Forum 2021 Dec 3 (Vol. 2, No. 12, pp. e214001-e214001). American Medical Association.
22. Barocas S, Selbst AD. Big data's disparate impact. Calif. L. Rev.. 2016;104:671.
23. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Annals of internal medicine. 2018 Dec 18;169(12):866-72.
24. Breiman L. Random forests. Machine learning. 2001 Oct;45(1):5-32.
25. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. Journal of chronic diseases. 1987 Jan 1;40(5):373-83.
26. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Medical care. 1998 Jan 1;36(1):8-27.
27. Schwartz AL, Zlaoui K, Foreman RP, Brennan TA, Newhouse JP. Health care utilization and spending in Medicare Advantage vs traditional Medicare: a difference-in-differences analysis. InJAMA Health Forum 2021 Dec 3 (Vol. 2, No. 12, pp. e214001-e214001). American Medical Association.
28. Schwartz AL, Zlaoui K, Foreman RP, Brennan TA, Newhouse JP. Health care utilization and spending in Medicare Advantage vs traditional Medicare: a difference-in-differences analysis. InJAMA Health Forum 2021 Dec 3 (Vol. 2, No. 12, pp. e214001-e214001). American Medical Association.
29. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Annals of internal medicine. 2018 Dec 18;169(12):866-72.
30. Breiman L. Random forests. Machine learning. 2001 Oct;45(1):5-32.

HOW TO CITE THIS ARTICLE: Gupta A. Predictive Risk Stratification in Medicare Advantage Using Interpretable Machine Learning: Reducing Cost While Improving Outcome Equity. *J Adv Sci Res*. 2025;16(08): 22-33 **DOI:** 10.55218/JASR.2025160804