Health Policy OPEN 10 (2026) 100160

Contents lists available at ScienceDirect

Health Policy OPEN

journal homepage: www.elsevier.com/locate/hpopen

Cross-sectional analysis of accuracy versus interpretability in Medicare Advantage risk adjustment

Maya Lozinski *

Department of Medicine, Massachusetts General Hospital, USA

ARTICLE INFO

Keywords:
Medicare risk adjustment
Accuracy
Complexity
Tabular machine learning
Upcoding incentives

ABSTRACT

Background: Risk adjustment models in Medicare Advantage determine annual payments of over $300 billion in public funds to private companies. Policymakers want risk adjustment models that are both accurate and interpretable to ensure appropriate use of public funds.

Methods: The trade-off between accuracy and interpretability from using standard machine learning (ML) models for risk adjustment was evaluated. A cross-sectional analysis was conducted using 2018–2019 Medicare claims with 3,602,618 beneficiaries. Multiple risk adjustment models were estimated, including traditional and ML-based approaches. Performance was assessed using out-of-sample mean absolute and squared error (MAE and MSE). Interpretability was measured using coefficient count and log-transformed coefficient count of models.

Results: ML models, especially gradient-boosted trees, significantly improved prediction accuracy relative to recent Medicare models, with MAE reductions of − 1,352 (95 % CI: −1,392, −1,316) and MSE reductions of − 5 (95 % CI: −9, −1). However, these improvements increased model complexity by more than 1000x and provided less than 0.1 % of the accuracy improvement per additional coefficient of a previous major model change. Notably, the predictions from gradient-boosted trees responded less to strategic diagnosis coding, reducing incentives to upcode.

Conclusions: Standard ML models can modestly improve predictive accuracy but substantially worsen model interpretability in risk adjustment. Future research is needed to improve accuracy in these models while maintaining the interpretability essential for oversight of public spending.

1. Introduction

Algorithmic predictions play an increasingly substantial role in allocating public healthcare funds. In the United States, Medicare Advantage risk adjustment formulas alone determine the distribution of over $300 billion annually in government payments to private companies, accounting for approximately 1 % of the U.S. gross domestic product. [1,2] These algorithms adjust payments to health insurers based on predicted patient costs, aiming to address adverse selection and ensure equitable healthcare access. However, these models must balance a fundamental tension between accuracy and interpretability.

1.1. Interpretability and accuracy in Medicare risk adjustment

Interpretability is crucial in Medicare Advantage risk adjustment, as the program allocates vast public funds to private companies. Given the financial implications, stakeholders closely monitor model parameters, down to minor coefficient adjustments. [3] Consequently, policymakers have historically prioritized transparent, interpretable models with strong clinical face validity. [4–6] This interpretability helps policymakers to identify issues with potential model decision rules, justify current regulations, and provide transparency into how substantial public funds are allocated.

However, existing risk adjustment techniques have been criticized for inaccuracies that lead to both overpayments and underpayments. These inaccuracies inflate Medicare costs and hinder healthcare for vulnerable populations. [6–11] Machine learning has emerged as a potential approach to improving risk adjustment accuracy. [9,13–16,18].

Unfortunately, machine learning models can also introduce substantial model complexity, which conflicts with policymakers' needs for interpretability. Limited attention has been given to the trade-off between model accuracy and interpretability in risk adjustment, partly because the field currently lacks clear metrics to measure model interpretability in a manner that aligns with policymakers' practical

* Corresponding author at: Resident Physician, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA.
E-mail address: mlozinski@mgh.harvard.edu.

https://doi.org/10.1016/j.hpopen.2025.100160
Received 3 April 2025; Received in revised form 22 October 2025; Accepted 15 December 2025
Available online 17 December 2025
2590-2296/© 2025 The Author. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

considerations. Without such a metric, it remains unclear whether the accuracy improvements provided by machine learning justify their potential reduction in interpretability.

1.2. Prior research

Existing research has examined model interpretability across various domains, [17] yet the implications for payment policy remain underexplored, despite their potential importance to healthcare policy. Prior studies have attempted to simplify risk adjustment by reducing the number of patient characteristics included in models. [12,15] In this work, model complexity is based on the number of patient characteristics included. For example, all models that use ten patient health conditions are equally complex, whether they are a linear regression with 10 coefficients or a random forest with hundreds of interaction terms. Other studies utilize machine learning to define more nuanced patient characteristics and incorporate novel functional forms, finding that their use enhances risk adjustment accuracy [18,19].

However, a critical gap remains in understanding the trade-off between accuracy and complexity if stakeholders evaluate not only the patient characteristics used but also how these characteristics interact to determine payments. This gap highlights the need to examine complexity from the standpoint of payment interpretability, which may yield new insights into the feasibility of machine learning approaches in risk adjustment policy.

More broadly, risk adjustment formulas are known to involve trade-offs between overall accuracy and other considerations, including subgroup accuracy, cost containment, and patient selection incentives. [9,20–23] However, the specific trade-off between accuracy and interpretability in risk adjustment remains poorly understood, despite its importance given the substantial public funds at stake.

1.3. Research Aims

This study empirically examines the trade-off between accuracy and payment interpretability in Medicare Advantage risk adjustment. To objectively measure this trade-off, we introduce a measure of model complexity tailored for payment policy, defined as the number of coefficients required to characterize all model predictions. Additionally, we take the logarithm of the number of coefficients to measure relative increases in complexity.

Then, using Medicare claims data, we compare multiple risk adjustment models—including both traditional approaches and machine learning algorithms—evaluating them based on out-of-sample accuracy and complexity. These potential improvements are benchmarked against previous large changes to Medicare risk adjustment to help determine whether the predicted accuracy gains justify implementation. Finally, we assess how these models influence incentives for strategic patient selection (“cream-skimming”) and diagnosis coding (“upcoding”).

2. Methods

2.1. Sample

A cross-sectional analysis was conducted using Medicare fee-for-service claims data from 2018 to 2019. The sample selection process closely followed Medicare’s standard approach for sample definition for risk adjustment model fitting. Medicare beneficiaries were included if they were aged $\geq$ 65 years, qualified for Medicare by age, and were enrolled in Medicare Part A and B for all of 2018 and at least 1 month of 2019. Individuals were excluded if they had end-stage renal disease, were enrolled in Medicaid or Medicare Advantage during the study period, or resided outside a U.S. state. However, unlike Medicare’s approach, this study did not exclude patients with a non-Medicare primary payer nor patients who were long-term institutionalized, due to data limitations.

The sample was divided into a training set (89 %) and a test set (11 %). Models were developed using the training set and evaluated using the test set.

2.2. Variables

Predictor variables included age, sex, and health conditions recorded in 2018. Health conditions were identified using diagnosis codes and mapped to Hierarchical Condition Categories (HCCs). The outcome variable was annualized healthcare spending in 2019.

All models utilized standard Medicare risk adjustment variables to ensure that performance differences were attributable to model structure rather than variable selection. For machine learning models, variables were formatted as single variables (e.g., age and sex) instead of as saturated indicators with pre-specified interactions, allowing the algorithms to identify optimal relationships between variables.

2.3. Model specifications

Three types of risk adjustment models were evaluated:

1. Standard Medicare Models: Three standard models were recalibrated using the study sample: a uniform payment model, Medicare’s Demographic model, and Medicare’s HCC model. The uniform payment model estimated the mean spending. The Demographic model incorporated 5-year age bins interacted with sex indicators in a linear regression framework. The HCC model implemented was the 2023 V24 HCC model for age-eligible, community-dwelling beneficiaries. This model included 86 hierarchical health condition indicators, seven health condition interactions, and health condition count indicator variables. [24]
2. Alternative Linear Models: New parametric risk adjustment models were developed using ordinary least squares and lasso regression. The new “HCCxCount” model incorporated interactions between health conditions and comorbidity counts to better account for the complexity of multimorbidity. Lasso regression models were employed to mitigate overfitting by eliminating coefficients with insufficient explanatory power. [25] The lasso models included both the HCC variables and the “HCCxCount” variables.
3. Non-Parametric Machine Learning Models: Tree-based models were developed, including regression trees, random forests, and gradient-boosted trees. Two versions of each approach were trained: one optimized to minimize mean squared error (MSE) and another to minimize mean absolute error (MAE). Models were implemented using Scikit-Learn and XGBoost. [26,27] Hyperparameters were optimized via random search with threefold cross-validation. Early stopping criteria were incorporated to limit the depth of trees in gradient-boosted trees. Additional details regarding the sample, variables, and model fitting are provided in the Appendix.

2.4. Measurement of model complexity

The interpretability of risk adjustment models was assessed based on the number of coefficients required to characterize model predictions.

For linear models, complexity was measured by the number of non-zero coefficients, also known as the L0 norm. For example, a linear model with an intercept and a coefficient for female would have a complexity of 2.

For tree-based regression models, complexity was quantified based on the number of unique, feasible decision paths that could be taken. In regression models with multiple trees, this measure equals the number of unique possible predictions that a model can make. For example, a simple regression tree that splits on sex and then, for men, on the presence of heart failure, has three possible predictions: one for women, one for men with heart failure, and one for men without heart failure.

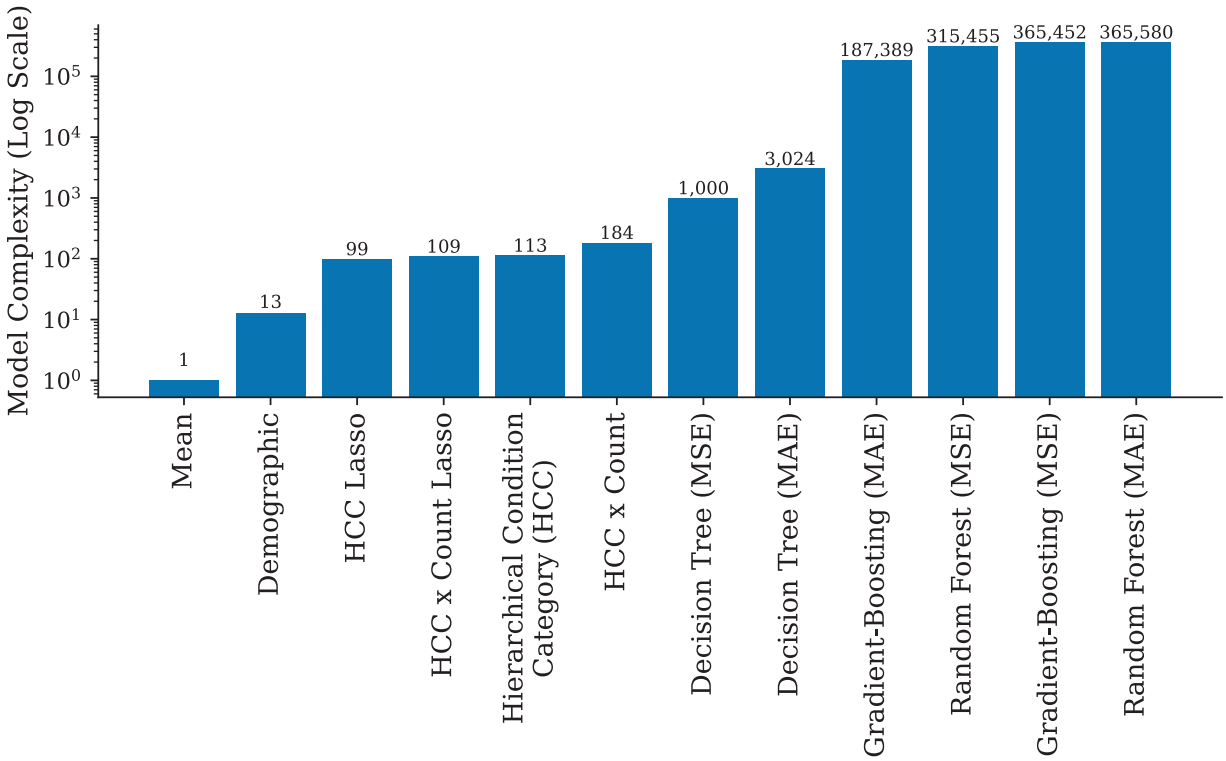


Fig. 1. Complexity (number of coefficients) of standard and machine learning models for risk adjustment. Notes: This graph shows the number of coefficients in selected risk adjustment models fit in the data. The x-axis shows the names of various models. The "Hierarchical Condition Category (HCC)" model is the current Medicare risk adjustment model. Models with "MAE" or "MSE" after their name indicate the objective function that the model was trained on, mean absolute error or mean squared error. The y-axis shows the count of coefficients, a proxy of the "complexity," or non-interpretability, of a model. For linear models, complexity is the number of non-zero coefficients, or the L0 norm. For tree-based models, complexity is the number of unique and feasible decision paths in the model. Note that the y-axis is in log scale, not linear scale, to display the full range of complexity values.

For computational feasibility, in models with multiple trees, complexity was estimated using the number of unique predictions *realized* in the training data, rather than the number *possible*, which provided a lower bound on the complexity of these already complex models. These measures reflect the complexity of generating predictions rather than fitting the models. This focus aligns with research suggesting interpretability should be evaluated based on domain-specific "cognitive chunks" that stakeholders process when understanding a model. [17,28] In risk adjustment, coefficients influence payment rates and are linked to interpretable clinical labels (e.g., heart failure), [24] making them cognitively salient and central to stakeholder interpretation. Notably, the original developers of the HCC model also assessed model complexity based on the number of coefficients, noting their new model used "fewer than 200 parameters." [4].

As an alternative measure, model complexity was quantified using the logarithm of the number of coefficients. This log-based measure captures relative, rather than absolute, increases in complexity. Increasing the coefficient count from 10 to 100 and from 100 to 1,000 both result in a one-unit increase in complexity, despite vastly different absolute changes. This approach assumes that the marginal complexity cost of adding the thousandth coefficient is lower than that of adding the hundredth coefficient.

2.5. Alternative complexity measures

Several alternative complexity measures were examined but ultimately not adopted due to their inability to reflect an essential intuition: a linear regression model with 100 indicator variables is simpler and more interpretable than a random forest model that uses the same variables but introduces extensive, non-linear interactions, resulting in thousands or even hundreds of thousands of unique predictions.

The first alternative measure considered was the number of input variables, as used in previous studies. [12,15] Under this definition, a linear model with 100 indicator variables would have the same complexity as a random forest model using the same input variables but with thousands of additional interaction terms. The second alternative was the number of unique potential predictions, which would classify a linear model with 100 indicator variables as having 2^{100} possible predictions, making it appear more complex than a random forest model using the same variables.

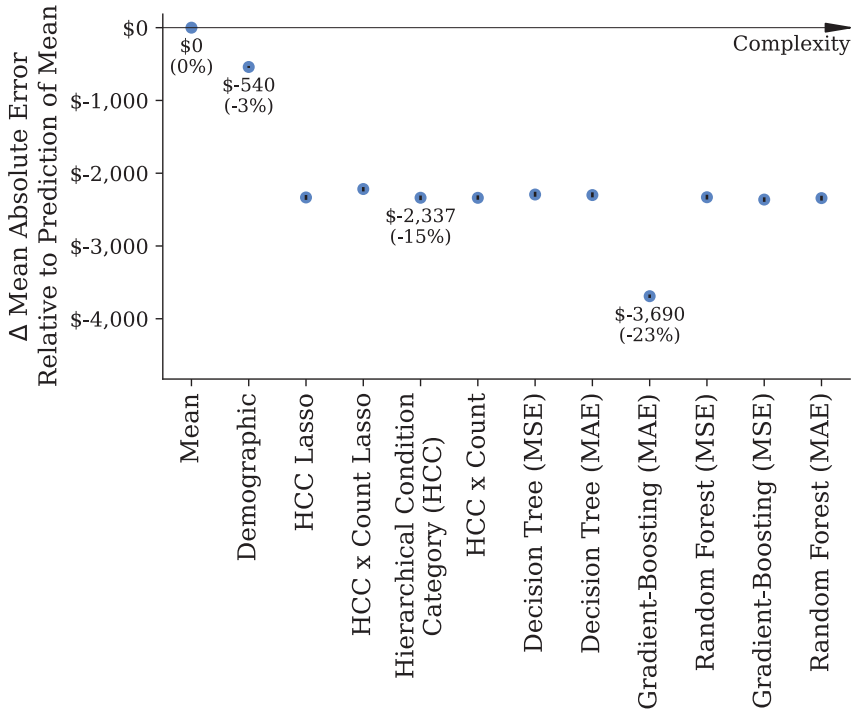
Notably, all complexity measures examined here assess global complexity (overall model complexity) rather than local complexity (the complexity of explaining individual payment decisions). [28] Unfortunately, defining comparable local complexity measures across model types is challenging, as the interpretation of coefficients depends on a model's overall structure.

2.6. Accuracy improvements

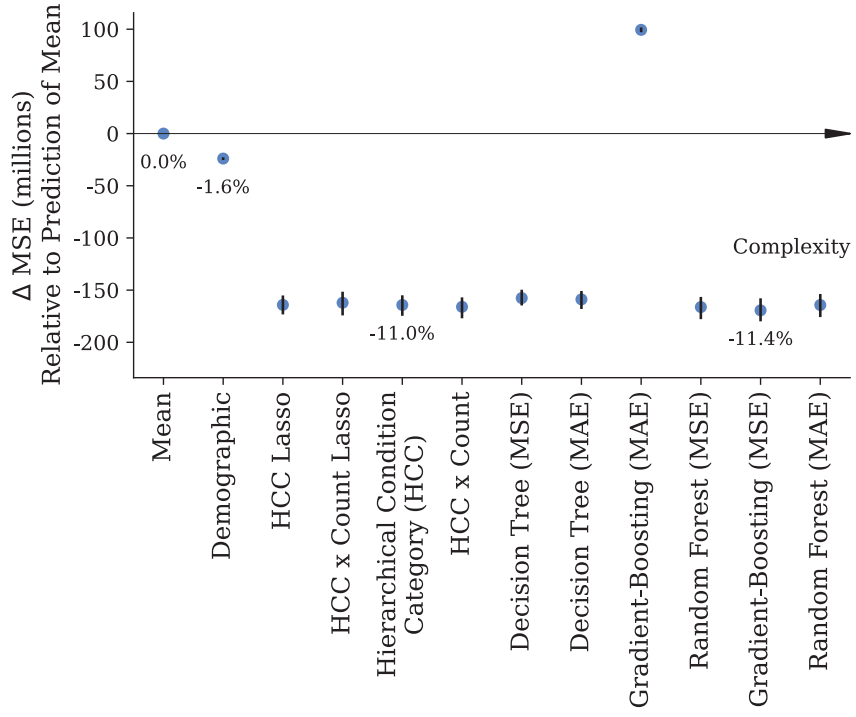
Model performance was evaluated using the mean absolute error (MAE) and the mean squared error (MSE) on the test dataset. Bootstrapping was employed to construct confidence intervals, following established methodology. [13] The MSE and MAE were calculated 100 times from bootstrapped datasets resampled with replacement at the patient level from the test dataset. The 2.5th and 97.5th percentile values of MAE and MAE were used as confidence bounds. Differences in MAE and MSE between models were also calculated in each bootstrapped dataset, creating confidence intervals for differences in performance between models.

To assess the trade-off between accuracy and complexity, we calculated the marginal improvements in both accuracy and complexity. These values were compared against historical changes in Medicare risk

(a) Difference in Mean Absolute Error of Model Predictions Relative to Predicting the Mean



(b) Difference in Mean Squared Error of Model Predictions Relative to Predicting the Mean



(caption on next page)

Fig. 2. Differences in Model Accuracy Relative to Predicting the Mean. Notes: These panels show the change in predictive performance, measured by mean absolute error (Panel (a)) and mean squared error (Panel (b)), for a variety of models relative to a baseline strategy of always predicting the mean. The x-axis lists the models, ordered by increasing complexity. The "Hierarchical Condition Category" model represents the current Medicare risk adjustment model. Models labeled "MAE" or "MSE" indicate the objective function used during training: mean absolute error or mean squared error, respectively. Both the MAE and MSE values are calculated out-of-sample using 10% of the available data. Each round marker denotes the point estimate of the performance improvement, and the black bars represent 95% confidence intervals estimated via bootstrapping. In many cases, these intervals are narrower than the height of the round markers. Values presented in parentheses show the percentage improvement relative to always predicting the mean.

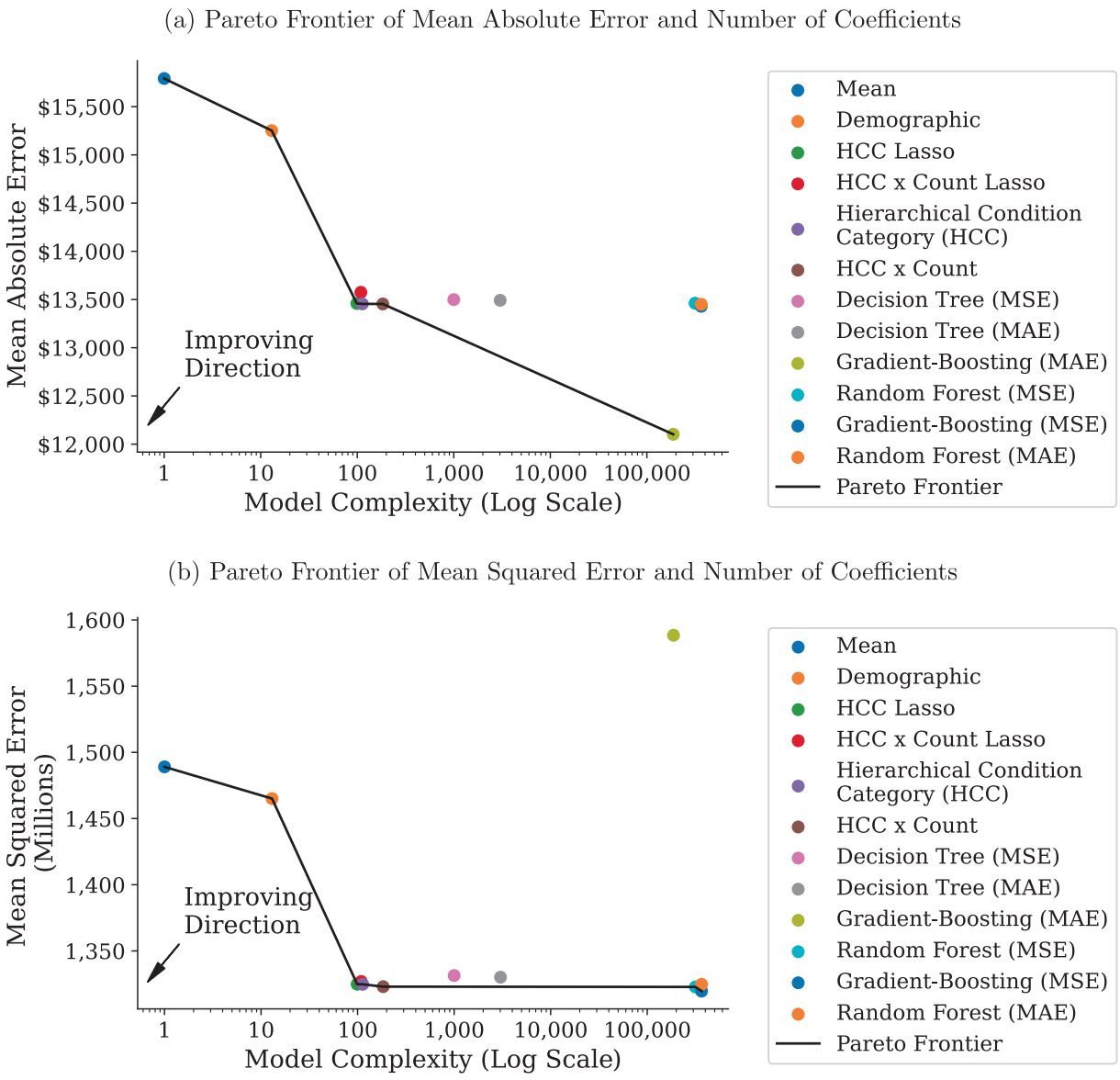


Fig. 3. Pareto Frontiers of Accuracy and Complexity. Notes: These panels show scatterplots of the accuracy and the number of coefficient in different risk adjustment models. Panel (a) displays the Mean Absolute Error (MAE) metric, while Panel (b) displays the Mean Squared Error (MSE) metric to measure accuracy. In both panels, the x-axis is plotted on a log scale and shows the number of coefficients. The y-axis shows model accuracy as measured by the respective metric (MAE or MSE) on out-of-sample predictions. The black line in each panel represents the Pareto frontier of accuracy and complexity. Improvements move toward the origin (i.e., lower complexity and higher accuracy).

adjustment models, specifically the transition from the Demographic model to the HCC model in 2004. [29] This major transition serves as a benchmark for acceptable complexity increases in exchange for improved model accuracy. More recent model revisions to the HCC model were also examined; however, these incremental changes led to noisy and non-significant improvements in performance, and therefore were of limited utility as benchmarks.

2.7. Gaming incentives

Models were evaluated on two key dimensions of gaming incentives: patient selection and upcoding. We examined whether the models reduced measures of selection incentives identified in prior literature. These measures were estimated overall and for patient subgroups based on chronic conditions known to be costly, [6,9] as defined in the Medicare Master Beneficiary Summary File. [30] The first measure of selection incentives was the sum of positive residuals, which indicates

Table 1
Summary of Risk Adjustment Model Performance—Complexity, Accuracy, and Marginal Improvements.

Model	Model Complexity (Number of Coefficients)	Out-of-sample MAE	Reduction in MAE vs. Uniform Payment	Out-of-sample MSE (millions)	Reduction in MSE (millions) vs. Uniform Payment	Pareto Model	MAE Change per Coef. from Prior Pareto Model	MSE Change per Coef. from Prior Pareto Model
Uniform Payment (Mean)	1	$15,792 [$15,683, $15,907]	$0	1,489 [1,184, 1,928]	0	X		
Demographic	13	$15,252 [$15,155, $15,370]	−$540 [$-554, $-525]	1,465 [1,186, 1,859]	−24 [-25, -23]	X	−$45/coef	−2.0/coef
HCC Lasso	99	$13,459 [$13,368, $13,566]	−$2,333 [$-2,366, $-2,300]	1,325 [1,049, 1,781]	−164 [-173, -155]	X	−$21/coef	−1.6/coef
HCC x Count Lasso	109	$13,575 [$13,463, $13,660]		1,327 [1,051, 1,745]				
Hierarchical Condition Category (HCC)	113	$13,455 [13,367, 13,555]		1,325 [1,058, 1,857]			−$0.29/coef	−0.0058/coef
HCC x Count	184	$13,454 [$13,353, $13,550]	−$2,338 [$-2,365, $-2,305]	1,323 [1,028, 1,813]	−166 [-177, -157]	X	−$0.012/coef	−0.025/coef
Decision Tree (MSE)	1,000	$13,499 [$13,411, $13,600]		1,331 [1,049, 1,730]				
Decision Tree (MAE)	3,024	$13,492 [$13,389, $13,595]		1,330 [1,054, 1,769]				
Gradient-Boosting (MAE)	187,389	$12,102 [$11,977, $12,225]	−$3,690 [$-3,714, $-3,667]	1,588 [1,304, 2,012]		X (MAE)	−$0.0072/coef	
Random Forest (MSE)	315,455	$13,462 [$13,366, $13,558]		1,323 [1,059, 1,830]	−166 [-178, -157]	X (MSE)		−0.00000069/coef
Gradient-Boosting (MSE)	365,452	$13,430 [$13,339, $13,521]		1,320 [1,041, 1,739]	−169 [-180, -158]	X (MSE)		−0.000063/coef
Random Forest (MAE)	365,580	$13,450 [$13,342, $13,551]		1,325 [1,057, 1,720]				

Notes: This table presents the complexity and predictive performance of the evaluated risk adjustment models. Column 1 lists the models, including the current Medicare risk adjustment model (Hierarchical Condition Category). Column 2 lists model complexity, which is measured by the number of coefficients: for linear models, this is the number of non-zero coefficients (L0 norm); for tree-based models, this is the number of unique and feasible decision paths. Columns 3 and 4 show MAE and MSE, measured out-of-sample, for each model. Columns 5 and 6 show the improvement relative to a baseline strategy of always predicting the mean (Uniform Payment). For columns 3-6, 95% confidence intervals are estimated via bootstrapping the out-of-sample data and are shown in brackets below the point estimate. In column 7, models marked with "X" in the Pareto Model column are on the Pareto frontier of accuracy and complexity. The final two columns show the marginal change in accuracy (MAE or MSE) per additional coefficient relative to the prior Pareto-optimal model.

the extent to which insurers have scope to select patients who cost less than predicted. [31] The second measure was expected net compensation, defined as the difference between average payments and costs. [9] The final two measures assessed the proportion of patients whose actual costs were either half or twice the predicted amount, representing insurers' exposure to positive and negative tail risks. [13].

Upcoding incentives were evaluated by measuring the increase in predicted spending resulting from the addition of specific diagnoses to patient records. The diagnoses were selected from industry promotional materials highlighting the "biggest HCC coding opportunities." [32].

3. Results

3.1. Sample characteristics

The study analyzed data from 3,602,618 Medicare beneficiaries, including 3,202,327 in the training sample and 400,291 in the test sample. Descriptive statistics for the training and test samples are provided in Appendix Tables 1 and 2. The average beneficiary in the training sample had 2.60 (SD = 3.56) payment-relevant health conditions and an average annualized spending of $13,449.85 ($SD$ = 35,441.55). The test sample exhibited similar characteristics.

3.2. Model complexity and accuracy

Figs. 1 and 2 show the number of coefficients and the accuracy of different models. The following discussion focuses on models along the Pareto frontier, where no other model offered better performance at lower complexity or vice versa. Fig. 3 presents the Pareto frontier of accuracy versus complexity. Table 1 summarizes these results.

In terms of complexity, the uniform payment model was the simplest, containing only one coefficient (mean spending). The current Medicare HCC model had moderate complexity, with 113 coefficients. In contrast, gradient-boosted trees optimized for MAE used 187,389 coefficients, while those optimized for MSE used 365,452 coefficients—three orders of magnitude more complex than the HCC model in absolute terms. When using log-transformed coefficient count, gradient-boosted trees increased model complexity by 2.6- and 2.7-fold for the MAE and MSE models, respectively.

Higher model accuracy required increased model complexity. The simplest model of uniform payment for each patient yielded a mean absolute error (MAE) of 15,792 (95 % CI: 15,683, 15,907) and a mean squared error (MSE) of 1,489 (95 % CI: 1,184, 1,928). The HCC model significantly improved accuracy, reducing MAE by − 2,337 (95 % CI: −2,370, −2,307) and MSE by −164 (95 % CI: −175, −155) relative to

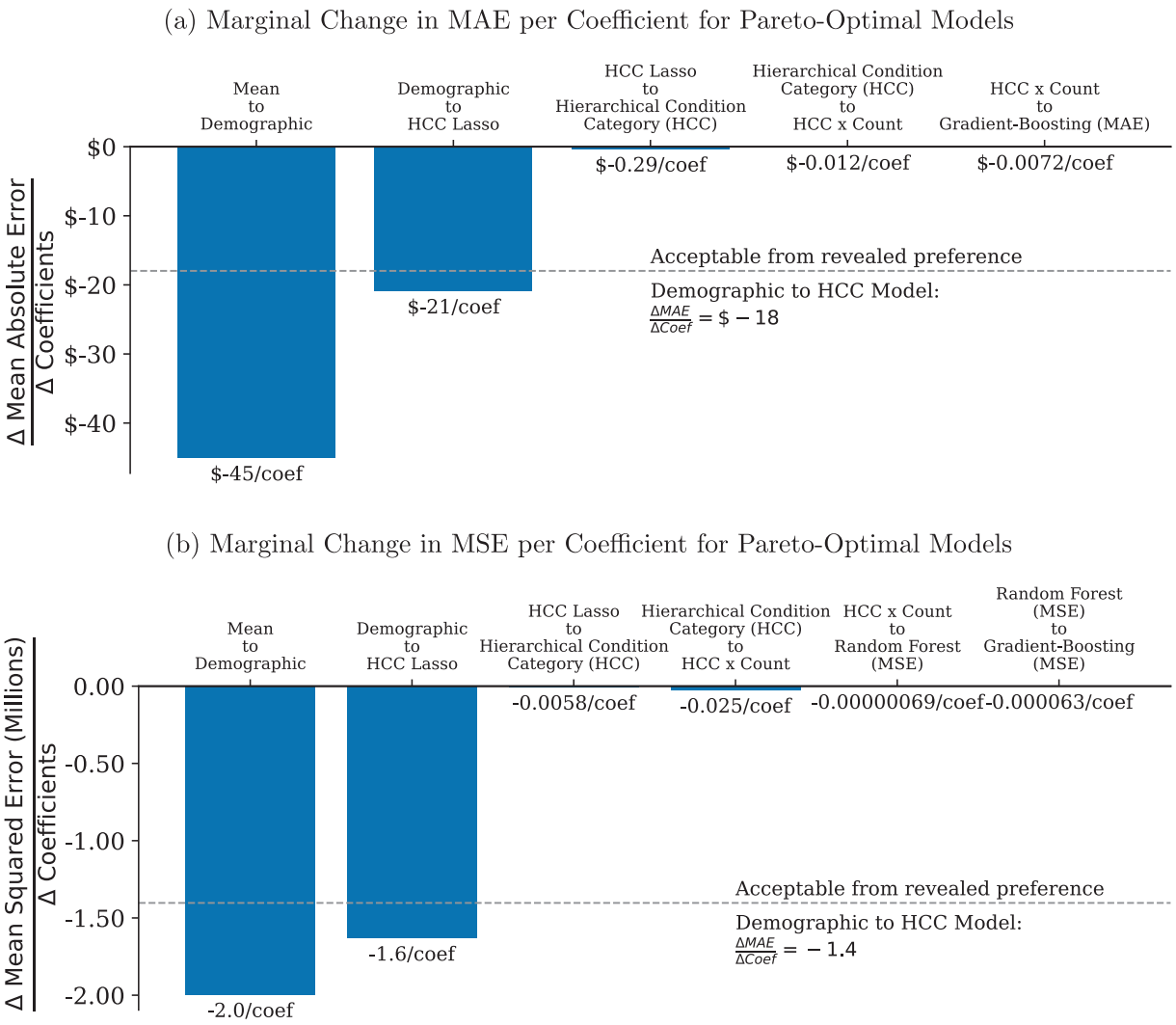


Fig. 4. Marginal Change in Accuracy per Coefficient for Models on the Pareto Frontier. Notes: These panels focus on models that are on the Pareto frontier, balancing accuracy and complexity. Panel (a) displays the marginal change in MAE per added coefficient, while Panel (b) displays the marginal change in MSE per added coefficient. In both panels, the x-axis lists model pairs in order of increasing complexity, and the y-axis shows the marginal improvement in the respective accuracy metric per additional model coefficient. The dotted horizontal gray line in each panel indicates the historical benchmark improvement (from the Demographic to the HCC model) that policymakers previously found acceptable.

uniform payment.

Gradient-boosted trees achieved the largest accuracy improvements. The MAE-optimized model reduced MAE by − 3,690 (95 % CI: −3,714, −3,667) compared with uniform payment and by − 1,352 (95 % CI: −1,392, −1,316) relative to the HCC model. The MSE-optimized model reduced the MSE by − 169 (95 % CI: −180, −158) compared with uniform payment and by − 5 (95 % CI: −9, −1) relative to the HCC model.

3.3. Value of additional complexity

To assess the trade-off between complexity and accuracy, model changes were benchmarked against historical adjustments in Medicare risk adjustment models, specifically the transition from the Demographic model to the HCC model. This transition reduced MAE by an estimated 17.97 per additional coefficient and MSE by an estimated 1.4e + 06 per additional coefficient.

Using the main complexity measure, the MAE-optimized gradient-boosted tree model reduced error by only an estimated 0.007 per additional coefficient compared with the next best model (HCCxCount model)—just 0.04 % of the improvement in accuracy per extra coefficient seen with the change from the Demographic to HCC model. The MSE-optimized gradient-boosted tree model was even less efficient, achieving only 0.004 % of the improvement seen in past changes. These results are shown in Fig. 4.

When applying log-based complexity measures, which account for relative increases in complexity, performance improved but remained below historical benchmarks. The MAE-optimized gradient-boosted tree model achieved only 23.5 % while the MSE-optimized model achieved only 32.9 % of the improvement per coefficient seen in past changes.

3.4. Effects on selection and upcoding incentives

Gradient-boosted trees produced mixed effects on selection incentives. These results are illustrated in Appendix Fig. 1. For example, the MAE-optimized model reduced the sum of positive residuals overall and among subgroups with costly chronic conditions relative to the HCC model. However, the MAE-optimized model increased negative tail risk (i.e., the probability that actual costs exceed twice the predicted amount), particularly for high-risk groups. This shift amplified disparities across subgroups and heightened selection incentives in certain cases. In contrast, the MSE-optimized model resulted in negligible

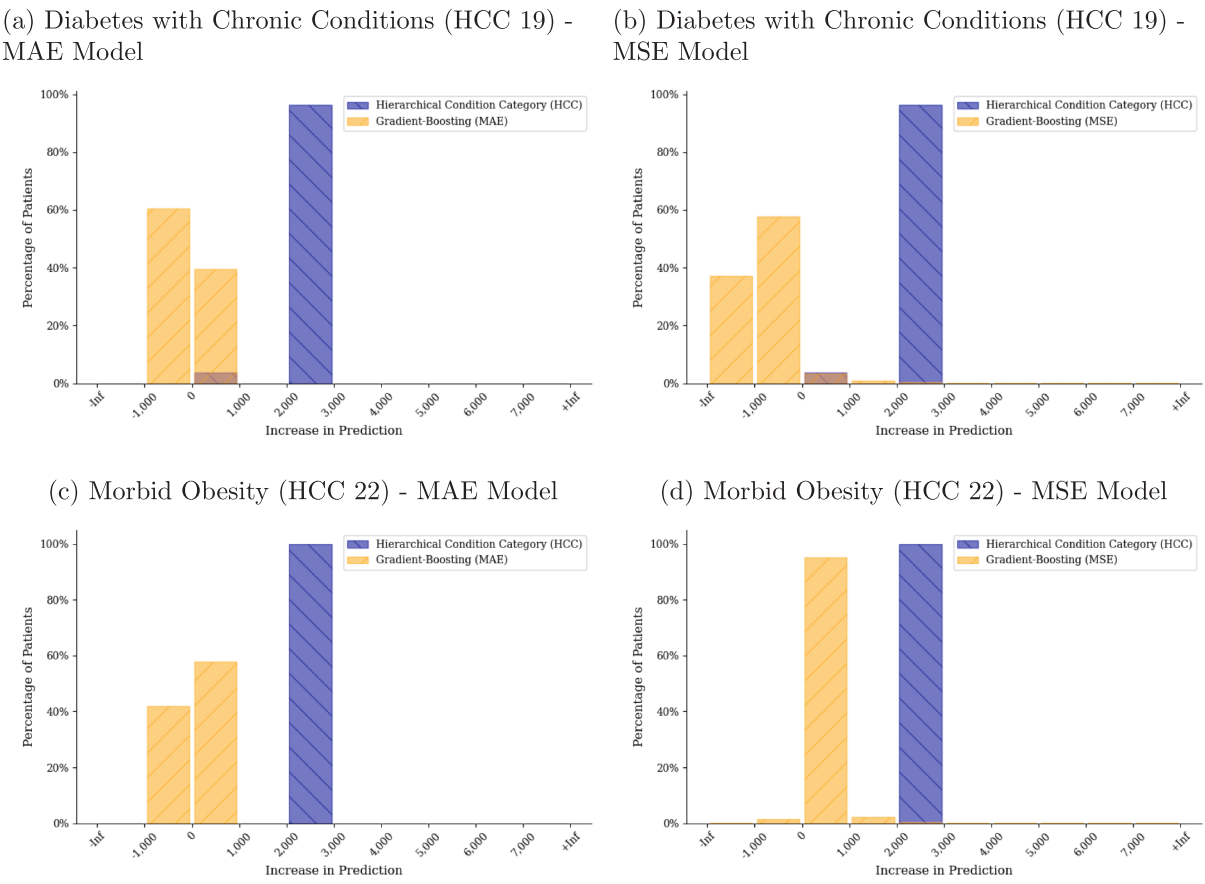


Fig. 5. Increase in Predicted Patient Cost from Upcoding Diabetes and Obesity (MAE and MSE Models). Notes: Panels (a) and (b) show the distribution of the change in predicted patient costs from adding diabetes with chronic complications (HCC 19) to patient records without HCC 19 for MAE and MSE models, respectively. If the patients have HCC 17 or HCC 18, milder types of diabetes, on their record, then those are set to zero. If they do not have HCC 17 or 18, their count of HCCs is increased by one, and any relevant diabetes interaction terms are set to one. Similarly, panels (c) and (d) show the same analysis for adding morbid obesity (HCC 22). The x-axis contains bins for the change in predicted spending. The y-axis shows the fraction of patients in the validation sample who fall into the bin. Bins with 10 or fewer individuals are suppressed per CMS requirements. Appendix Figs. 7 and 8 show the results for other chronic conditions.

changes to selection incentives.

Regarding upcoding incentives, gradient-boosted trees consistently reduced the financial gains from adding diagnoses to patient records. Fig. 5 presents the predicted spending increases associated with upcoding diabetes with chronic complications and morbid obesity, the two most prominent upcoding opportunities identified in industry materials. [32] In both cases, the HCC model produced greater increases in predicted spending compared with gradient-boosted trees. Figs. A3 and A4 confirm similar patterns for other chronic conditions identified as high-priority upcoding opportunities.

4. Discussion

4.1. Summary and interpretation

This study empirically assessed the trade-off between accuracy and interpretability in Medicare Advantage risk adjustment models. Machine learning models, particularly gradient-boosted trees, demonstrated higher predictive accuracy compared with a recent HCC model. However, these improvements came at the cost of substantially increased model complexity.

The effects on gaming incentives were mixed but intriguing. The effects on selection incentives were variable. However, notably, gradient-boosted trees improved prediction stability against strategic diagnosis coding, thereby reducing potential financial incentives for upcoding.

4.2. Interpretation

These findings suggest that standard machine learning models, despite their superior accuracy, are unlikely to provide viable solutions for improving Medicare Advantage risk adjustment. The complexity required to achieve even modest accuracy gains is beyond what policymakers would likely consider acceptable, especially if their preferences have remained unchanged over time. As such, standard off-the-shelf models are not well-suited for implementation in Medicare risk adjustment, given the need for model interpretability and transparency.

However, newer models could complement current CMS strategies for risk adjustment. Notably, gradient-boosted tree models were more resilient to upcoding than standard models, an intriguing finding that suggests machine models may have the potential to mitigate this widespread and costly strategic coding behavior. [34] In addition, machine learning models could be used to create new, more informative variables, discover interactions that are incompletely accounted for in current models, and improve current models in other ways.

4.3. Strengths

This evaluation has several key strengths. First, this work introduces a formal measure of model complexity informed by the interpretability literature and tailored for healthcare payment policy. This measure enables quantitative assessment of the tradeoff between accuracy and interpretability across a wide range of contemporary machine learning

models. In addition, this assessment is performed on a sample of claims from over 3 million Traditional Medicare beneficiaries, which limits small sample variation and strengthens the external validity.

4.4. Limitations

This study has several key limitations. First, the analysis closely but not identically followed Medicare risk adjustment sample selection and model fitting due to data limitations. For example, the historical Demographic model was fit on 2018 data. In addition, the findings may not generalize beyond the 2018–2019 period studied or to other payer contexts, such as Medicaid or employer-sponsored insurance.

Second, only a subset of potential functional forms was tested, although the study included leading machine learning models for regression problems [33].

Third, the measure of model complexity was derived from the interpretability literature and secondary policy document analysis, but it has not been formally validated with policymakers. Future research involving human subjects is needed to definitively assess how policymakers and healthcare administrators cognitively process model complexity in payment settings.

Fourth, improvements in risk adjustment were benchmarked against changes implemented in 2004, which may not align with current policymaker preferences, particularly if they have changed over time.

Fifth, the study assessed upcoding robustness by introducing one diagnosis at a time for select conditions, which may not fully capture real-world coding practices.

Finally, the research focused on a single trade-off: accuracy versus interpretability. Other considerations are also critical in the design of risk adjustment, including equity, subgroup accuracy, cost containment, and data collection burden, which are not considered here.

4.5. Future directions

The findings presented here have implications that extend beyond Medicare Advantage. Given that approximately two-thirds of Medicare and Medicaid spending is allocated through risk-adjusted payments, [35,36] similar accuracy-interpretability trade-offs likely exist across healthcare payment policies.

These results highlight a broader challenge in healthcare policy: mathematically optimal solutions may not be policy optimal when governance constraints are considered. [37] Future work should prioritize the development of constrained machine learning approaches tailored to policy contexts, rather than retrofitting standard techniques. Recent work in risk adjustment has pioneered in this direction. [18,19] Our study emphasizes the significance of this recent work and the necessity for future efforts to create more tailored models that improve both accuracy and interpretability across various risk adjustment contexts.

5. Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used Claude and ChatGPT to help summarize the research and reduce word counts. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

CRediT authorship contribution statement

Maya Lozinski: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The author declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The author thanks David Meltzer, Dan Black, Tamara Konetzka, Zarek Brot-Goldberg and Joshua Gottlieb for advising, feedback and support with data access, and Aaron Schwartz, Ari Anisfeld, and Anna Zink for feedback and comments. Thank you to the following agencies for funding: Agency for Healthcare Research and Quality (R36HS028592), NIH Medical Scientist Training Program (T32GM007281), NIA Program in Medicine, the Social Sciences, and Aging (T32AG051146), and the Center for Research Informatics (2U54TR002389-06).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.hpopen.2025.100160.

References

[1] Statista Research Department. U.S. Gross Domestic Product: Forecast 2021-2032. Published 2022. Accessed May 25, 2023.
[2] Cubanski J, Neuman T. The Facts on Medicare Spending and Financing. Henry J: Kaiser Family Foundation; 2023.
[3] James M, Stearns M, Rykaczewski K. A First Look at the 2024 CMS Advance Notice. Wolters Kluwer. Published 2023. Accessed June 7, 2023. https://www.wolterskluwer.com/en/expert-insights/a-first-look-at-the-2024-cms-advance-notice.
[4] Pope GC, Kautter J, Ellis RP, et al. Risk adjustment of Medicare capitation payments using the CMS-HCC model. Health Care Financ Rev. 2004;25:119.
[5] Medicare Payment Advisory Commission. Report to the Congress: Medicare and the Health Care Delivery System. Published 2014.
[6] Medicare Payment Advisory Commission. Report to the Congress: Risk Adjustment in Medicare Advantage. Published 2021.
[7] Geruso M, Layton TJ. Selection in health insurance markets and its policy remedies. J Econ Perspect. 2017;31(4):23–50.
[8] Rose S, Bergquist SL, Layton TJ. Computational health economics for identification of unprofitable health care enrollees. Biostatistics. 2017;18(4):682–94.
[9] Zink A, Rose S. Fair regression for health care spending. Biometrics. 2020;76(3):973–82.
[10] Zhu JM, Meiselbach MK, Drake C, Polsky D. Psychiatrist networks in Medicare Advantage plans are substantially narrower than in Medicaid and ACA markets. Health Aff (Millwood). 2023;42(6):909–18. https://doi.org/10.1377/hlthaff.2023.00163.
[11] Ryan AM, Chopra Z, Meyers DJ, Fuse Brown EC, Murray RC, Williams TC. Favorable selection in Medicare Advantage is linked to inflated benchmarks and billions in overpayments to plans. Health Aff (Millwood). 2023;42(8):1190–7. https://doi.org/10.1377/hlthaff.2023.00249.
[12] Rose S. A machine learning framework for plan payment risk adjustment. Health Serv Res. 2016;51(6):2358–74.
[13] Park S, Basu A. Alternative evaluation metrics for risk adjustment methods. Health Econ. 2018;27(3):e1–5.
[14] Kan HJ, Kharrazi H, Chang HY, et al. Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. PLoS One. 2019;14(3):e0213258. https://doi.org/10.1371/journal.pone.0213258.
[15] McGuire TG, Zink AL, Rose S. Improving the performance of risk adjustment systems: constrained regressions, reinsurance, and variable selection. Am J Health Econ. 2021;7(4):497–521.
[16] Irvin JA, Kondrich AA, Ko M, et al. Incorporating machine learning and social determinants of health indicators into prospective risk adjustment for health plan payments. BMC Public Health. 2020;20(1):1–10.
[17] Rudin C. Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15.
[18] Ellis RP, Hsu HE, Siracuse JJ, et al. Development and Assessment of a New Framework for Disease Surveillance, Prediction, and Risk Adjustment: The Diagnostic Items Classification System. JAMA Health Forum. 2022;3(3):e220276. https://doi.org/10.1001/jamahealthforum.2022.0276.
[19] Andriola C, Ellis RP, Siracuse JJ, et al. A Novel Machine Learning Algorithm for Creating Risk-Adjusted Payment Formulas. JAMA Health Forum. 2024;5(4):e240625. https://doi.org/10.1001/jamahealthforum.2024.0625.
[20] Ellis RP, McGuire TG. Predictability and predictiveness in health care spending. J Health Econ. 2007;26(1):25–48.

[21] Geruso M, McGuire TG. Trade-offs in the design of health plan payment systems: Fit, power, and balance. J Health Econ. 2016;47:1–19.
[22] Layton TJ, McGuire TG, Van Kleef RC. Deriving risk adjustment payment weights to maximize efficiency of health insurance markets. J Health Econ. 2018;61:93–110.
[23] Ellis RP, Martins B, Rose S. Risk adjustment for health plan payment. Risk Adjustment, Risk Sharing, and Premium Regulation in Health Insurance Markets. Springer; 2018.
[24] Centers for Medicare & Medicaid Services. CMS-HCC Software V24. Published 2023. Accessed April 28, 2023. https://www.cms.gov/files/zip/2023-initial-icd-10-mappings.zip.
[25] Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer; 2009.
[26] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
[27] Chen T, Guestrin C. In: XGBoost: A scalable tree boosting system. New York, NY: ACM; 2016. p. 785–94.
[28] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. 2017.
[29] McGuire TG, Newhouse JP, Sinaiko AD. An economic history of Medicare Part C. Milbank Q. 2011;89(2):289–332.
[30] MBSF CHRONIC 2023. Published 2023.
[31] Brown J, Duggan M, Kuziemko I, Woolston W. How does risk selection respond to risk adjustment? New evidence from the Medicare Advantage program. Am Econ Rev. 2014;104(10):3335–64.
[32] RCX Rules. An Overview of Medicare. Published 2023. Accessed September 25, 2023.
[33] Shwartz-Ziv R, Armon A. Tabular data: Deep learning is not all you need. Inf Fusion. 2022;81:84–90.
[34] Geruso M, Layton T. Upcoding: Evidence from Medicare on squishy risk adjustment. J Polit Econ. 2020;128(3):984–1026.
[35] Medicare Trustees. 2022 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds. Published 2022.
[36] Kaiser Family Foundation (KFF). Total Medicaid MCO Spending. Published 2023. Accessed May 25, 2023.
[37] Weissman GE, Joynt Maddox KE. Guiding Risk Adjustment Models Toward Machine Learning Methods. JAMA. 2023;330(9):807–8. https://doi.org/10.1001/jama.2023.12920.