Accuracy and Interpretability in Government Payment Algorithms*

Maya Lozinski
University of Chicago

October 2023

Abstract

I empirically investigate the trade-off between accuracy and interpretability in Medicare Advantage risk adjustment models. I introduce a formal metric for model complexity in payment policy, which equates complexity to the number of coefficients in a model, a factor central to stakeholder interpretation of payment rates. Machine learning models significantly improve prediction accuracy and robustness to upcoding but also dramatically increase complexity. An analysis of policymakers' preferences reveals that these models likely do not justify their additional complexity. Future research should explore aligning machine learning advances with payment policy constraints.

Keywords: healthcare, risk adjustment, machine learning
JEL codes: I18, C14

*Thank you to David Meltzer, Dan Black, Tamara Konetzka, Zarek Brot-Goldberg and Joshua Gottlieb for advising, feedback and support with data access. Thank you to Aaron Schwartz, Ari Anisfeld, and Anna Zink, as well as the ASHEcon and Harris PhD Workshop participants for feedback and comments. Thanks to the Agency for Healthcare Research and Quality (R36HS028592, PI: Maya Lozinski). In addition, thanks to the NIH Medical Scientist Training Program (T32GM007281), the NIA Program in Medicine, the Social Sciences, and Aging (T32AG051146), and the Center for Research Informatics (2U54TR002389-06). Author's email address: mayalozinski@uchicago.edu.

Accuracy and Interpretability in Government Payment Algorithms

Abstract: I empirically investigate the trade-off between accuracy and interpretability in Medicare Advantage risk adjustment models. I introduce a formal metric for model complexity in payment policy, which equates complexity to the number of coefficients in a model, a factor central to stakeholder interpretation of payment rates. Machine learning models significantly improve prediction accuracy and robustness to upcoding but also dramatically increase complexity. An analysis of policymakers' preferences reveals that these models likely do not justify their additional complexity. Future research should explore aligning machine learning advances with payment policy constraints.

1 Introduction

Algorithmic predictions are used in government decision-making to determine the flow of billions of public dollars. These algorithms commonly rely on regression models to predict values such as property values for tax purposes or a patient's annual total healthcare expenditures in public health insurance. Policymakers often face a trade-off between using simple, transparent statistical models and achieving higher predictive accuracy. While machine learning techniques offer potentially increased prediction accuracy, they often require policymakers to use complex non-parametric models.

Medicare Advantage exemplifies this tension. In Medicare Advantage, risk adjustment modifies payment amounts to health insurers based on expected patient costs, in order to address the "market for lemons", or adverse selection. In 2021, its risk adjustment formulas determined the allocation of over $300 billion, or 1% of US GDP (Statista Research Department, 2022; Cubanski and Neuman, 2023). Given the program's scale, even small changes to these formulas can change the allocation of billions in public funds.

Given the vast public funds involved, simple and interpretable risk adjustment is a key policymaker objective. For example, during the first 15 years of Medicare Advantage, Medicare used a model with only a few demographic variables, known as the "Demographic Model". This model was used even though it explained only 1% of the variation in spending and evidence existed that it contributed to selection, substantially increasing Medicare's costs. The model was made more complex only when congressional legislation required that it add adjustments for health status (McGuire, Newhouse, and Sinaiko, 2011). In response, Medicare added diagnosis information to risk adjustment and created the current Hierarchical Condition Category (HCC) model. Although this model added complexity, the developers of the HCC model still emphasized that preserving interpretability and parsimony were central design of the model (Pope et al., 2004). More recently, in reports to Congress, Medicare has emphasized that risk adjustment models should be transparent, interpretable, and clinically meaningful with "face-validity" (MedPAC, 2014, 2021), likely because this makes them more

defensible in the face of public scrutiny.

However, the program's current risk adjustment techniques have been criticized for their inaccuracy, leading to over- and under-payments (Rose, Bergquist, and Layton, 2017; Zink and Rose, 2020; MedPAC, 2021). Such payment discrepancies distort market dynamics and compromise healthcare access for vulnerable populations (Geruso and Layton, 2017). The HCC model and associated policy reforms appear to have mitigated selection (Newhouse et al., 2015) but the remaining selection continues to be substantial, costly, and harmful to patients (Ryan et al., 2023; Zhu et al., 2023). A 2017 review concluded that despite theoretical and practical challenges, improvements to risk adjustment offer "the best tool we have to address selection across plans in competitive health insurance markets" (Geruso and Layton, 2017).

Machine learning provides an avenue toward more accurate risk adjustment to address selection. However, machine learning models are frequently complex and hard to understand, causing them to fall short of stated yet mathematically informal interpretability requirements.

In light of these challenges, this paper aims to empirically examine the trade-off between the accuracy and interpretability of risk adjustment models in Medicare Advantage. Specifically, I evaluate whether machine learning models, despite their complexity, offer improvements in prediction accuracy that justify their use.

To measure this trade-off, I introduce a formal measure of model complexity tailored for government payment policy. I argue that model complexity (or lack of interpretability) is best proxied in this context as the number of coefficients in a model. This proxy assumes that the number of coefficients reflects the number of objects, or "cognitive chunks," that stakeholders consider when interpreting a risk adjustment model.

Using Medicare claims data, I fit and evaluate multiple risk adjustment models. These models include both conventional and machine learning models, and they are fitted on Medicare claims data using standard risk adjustment variables. The models use the same un-

derlying variables and differ solely in their functional forms. For each model, I estimate its complexity and out-of-sample accuracy.

The results show the clear trade-off between model accuracy and complexity. I find that the non-linear models provide the largest improvements in accuracy. A gradient-boosted tree changes mean absolute error (MAE) by $-1,352 (CI: $-1,392, $-1,316) relative to the current Medicare model, the HCC model. I also find that predictions from the gradient-boosted tree are more stable in the presence of simulated upcoding than predictions from the HCC model. However, this model also increases complexity by orders of magnitude, to 187,389, from 113 in the current HCC model.

To assess whether this increased complexity is justifiable, I use past changes in Medicare risk adjustment to estimate a range of plausible preferences of accuracy relative to complexity. I find that policymakers have accepted model changes that reduce MAE by $17.97 per additional coefficient. New models provide a reduction of $0.00722 per additional coefficient. As such, for new models to be acceptable, policymakers would need to be willing to accept model changes that are considerably less efficient at reducing error than they have in the past. These findings are robust to alternative accuracy metrics and preference assumptions about the disutility of complexity. They are limited to the extent that greater accuracy reduces selection incentives. As a whole, these findings suggest that standard machine learning models alone are unlikely to provide acceptable solutions to current issues with risk adjustment accuracy.

This work has several contributions. First, it provides a direct comparison between current and proposed risk adjustment models for Medicare Advantage. While prior research has suggested the potential for machine learning models to improve the accuracy of risk adjustment in Medicare Advantage (Rose, 2016; Park and Basu, 2018; Kan et al., 2019; McGuire, Zink, and Rose, 2020; Zink and Rose, 2020; Irvin et al., 2020), these studies rely on different datasets, typically commercial claims data that cover a younger, healthier population. They also often focus on risk adjustment in other settings, such as ACA exchanges. Given that

individuals on Medicare exhibit more complex patterns of health conditions and spending, the value-add of machine learning is expected to be higher in this setting. Therefore, the conclusion—that machine learning is not worth the additional complexity—is more compelling, as it comes from a direct comparison with standard Medicare Advantage models where the value-add is expected to be larger.

This paper also contributes to the literature identifying trade-offs in the design of health insurance risk adjustment formulas. Ellis and McGuire (2007) observe a trade-off between "fit" (accuracy), "power," and "balance" generated by different risk adjustment formulas. Zink and Rose (2020) observe a trade-off between global accuracy and accuracy for certain patient subgroups in risk adjustment. Layton, McGuire, and Van Kleef (2018) argue that risk adjustment should maximize a welfare-grounded objective function, rather than R^2, highlighting a trade-off between accuracy and welfare. This paper contributes by identifying a key trade-off between accuracy and interpretability, which arises due to governance constraints on risk adjustment.

This study also has broader implications for policy settings in the US and internationally. Risk adjustment is used widely in US healthcare policy, with roughly two-thirds of Medicare and Medicaid dollars allocated via risk-adjusted payments, totaling over $1 trillion annually (Medicare Trustees, 2022; KFF, 2023). Other countries, including Germany, Netherlands, Switzerland, and Chile (Kautter, Pope, and Keenan, 2014; Henriquez et al., 2023), also employ risk adjustment in their publicly regulated health insurance markets. Payment formulas are also used to assess property taxes in the US and worldwide (Norregaard, 2013; Berry, 2021). Hence, the trade-off between accuracy and complexity is broadly relevant, and these findings can inform efforts to incorporate machine learning into payment policy in those settings as well.

This work also introduces a new domain application to the machine learning interpretability literature. Existing research has considered model interpretability across numerous domains (Rudin, 2019) but payment policy remains relatively unexplored despite its growing

importance in the US and other countries. Rose (2016) and McGuire, Zink, and Rose (2020) consider how to simplify risk adjustment, but their motivation is to reduce opportunities for gaming, not to improve interpretability. They arrive at a different measure, which leads to substantially different conclusions about the complexity and policy feasibility of machine learning models. More broadly, this paper highlights that interpretability is a key barrier to using machine learning in payment policy and identifies it as an important domain for future research.

2 Complexity in risk adjustment

2.1 Risk adjustment accuracy

The goal of risk adjustment is not perfect accuracy but rather to maximize accuracy conditional on using only "appropriate" variation. Broadly, appropriate variation is variation that contributes to selection incentives (i.e., higher expected costs) but does not lead to manipulation or moral hazard by insurance companies. This distinction is operationalized, albeit imperfectly, by including only variables with "appropriate" variation in risk adjustment, such as demographics and health conditions, and excluding ones with inappropriate variation, like past spending (Geruso and Layton, 2017).

2.2 The need for interpretability

Maintaining model interpretability and face validity is also a policy priority. Risk adjustment rate setting controls the flows of vast amounts of public funds to private companies, so these models are subject to intense public scrutiny. Medicare publicly releases all model parameters, a contrast to other policy algorithms, such as bail decisions, where parameters are often proprietary (Rudin, 2019). Tweaks in models are closely followed by trade press

and private companies, down to small changes in coefficient values.[1] As such, this need for transparent, face-valid models appears to stem from political and governance constraints.

Medicare risk adjustment also has additional requirements, or what the machine learning literature refers to as "auxiliary criteria." For example, per reports to Congress, risk adjustment must have face validity, a criterion I interpret as having two parts. First, a model should be interpretable, and second, upon interpretation, it must pass unspecified "sniff tests." These auxiliary criteria are not fully formalized and therefore are not included in the objective function. Interpretability allows the Centers for Medicare and Medicaid Services (CMS) to ex-post assess—and, if needed, to enforce—these auxiliary criteria that have not been included in the objective function (Doshi-Velez and Kim, 2017).

The presence of auxiliary criteria also explains why a simple rule like "minimize mean squared error" is an insufficient criterion by which to judge alternative risk adjustment models. Narrowly focusing on MSE makes one liable to generate models that do not meet necessary auxiliary criteria and are therefore acceptable to policymakers.

Of note, I must clarify what impact, if any, interpretability has on gaming incentives. Interpretability will likely worsen selection incentives if it necessitates simpler, less accurate models, as discussed above. However, it is theoretically ambiguous how complex, non-linear models change the incentives to upcode.

2.3 Measure of model complexity

This study introduces a precise measure of model complexity, defined here as non-interpretability, for payment policy contexts: the number of coefficients in a model. Simply put, the model asks how many parameters are necessary to generate the full range of predictions.

For linear models, this measure is the L0 norm, or the number of non-zero coefficients. For example, a linear model with an intercept and a coefficient for female would have a

[1] To quote one recent trade press article: "How will ... the changes in the proposed coefficients financially impact your organization? ... The time to act is now! CMS will be accepting commentary through Friday, March 3, 2023" (James, Stearns, and Rykaczewski, 2023).

complexity of two.

For tree-based models, this measure of complexity is determined by the number of unique, feasible decision paths, or equivalently, the number of combinations of variable values that lead to distinct predictions. For example, consider a regression tree that splits only sex and then, for men only, the presence of heart failure. This model can be represented with three coefficients: one for women, one for men with heart failure, and one for men without heart failure. Each coefficient represents the predicted spending for each group, giving the model a complexity of three.

The approach to calculating complexity differs subtly for tree-based models like gradient-boosted trees and random forests, which are both collections of trees where the predictions of each tree are combined in sums or averages. Consider a second tree that splits only on sex and then, for women only, on the presence of diabetes. This second tree also has an individual complexity of three. However, only four coefficients are needed to express the predictions made by combining trees, one for each of the following groups: men with and without heart failure and women with and without diabetes. As such, the combined trees have a complexity of four.

2.4 Strengths and limitations of the complexity measure

Why is the number of coefficients an appropriate measure of model complexity (or non-interpretability) in this context? The machine learning interpretability literature argues that interpretability should be considered through an explanation's basic units or "cognitive chunks" (Doshi-Velez and Kim, 2017), which are domain specific (Rudin, 2019; Doshi-Velez and Kim, 2017). These cognitive chunks reflect the complexity of explaining a model, not fitting it. In risk adjustment, coefficients vary payment rates, which makes them likely to be highly cognitively salient to stakeholders. In addition, coefficients have interpretable labels, e.g., heart failure (CMS, 2023), suggesting that stakeholders are inspecting and interpreting the model at this level. Lastly, the HCC model developers used the number of coefficients

as an informal measure of model complexity, observing that the HCC model was relatively parsimonious at "fewer than 200 parameters" (Pope et al., 2004).

There are two key limitations. First, the evidence for this measure is based on secondary interpretation of policy documents and papers. Definitive assessment requires human subjects research (Doshi-Velez and Kim, 2017), which is outside this paper's scope.

Second, this definition is a measure of *global* complexity, the model's overall complexity, to use parlance from Doshi-Velez and Kim (2017). It does not directly measure local complexity, the complexity of explaining individual payment decisions. A measure of local complexity would be desirable given that policymakers often need to justify individual decisions. Unfortunately, it is difficult to pin down reasonable and comparable measures of local complexity across models because the interpretation of any set of coefficients depends on the structure of the rest of the model.

2.5 Alternative measures of model complexity

An alternative measure of global model complexity is the number of substantive input variables used in a model, as in Rose (2016); McGuire, Zink, and Rose (2020). Under this definition, a linear model with 100 input variables would be considered equally interpretable as a random forest with the same 100 input variables and hundreds of thousands of interaction terms. This is because this definition does not count interaction terms in the salient cognitive chunks. I argue that this is an incomplete assessment of what matters to stakeholders in risk adjustment. Interaction terms impact payments, so stakeholders are likely to demand explanations that account for them. This makes such terms highly relevant cognitive chunks when assessing interpretability.

Definitive arbitration of the saliency of interaction terms would require human subject experiments (Doshi-Velez and Kim, 2017). But in the absence of such experiments, circumstantial evidence will have to suffice. The original developers of the HCC model specifically considered the clinical face-validity of interaction terms, implying they believe interactions

are subject to interpretation and scrutiny (Pope et al., 2004).

One other measure of global model complexity from a related literature is the number of unique potential predictions of a model (Kleinberg and Mullainathan, 2019). This definition argues the salient object is the prediction itself and that the process used to arrive at the prediction is irrelevant. According to this definition, a linear model with 100 indicator variables has a complexity of 1.3×10^{30} (2^{100}, or one nonillion possible unique predictions). A fully saturated tree-based model will have the same number of possible predictions, though most tree-based models will have fewer. As such, the linear model is weakly *more* complex than the random forest. This definition of complexity, when applied to the models used by Medicare policymakers, suggests that Medicare policymakers are currently using a maximally complex model. Therefore, they have no distaste for complexity whatsoever, rendering this exercise unnecessary.

Why not just try to explain non-interpretable models rather than require interpretable models? Simplified explanations of complex models are necessarily inaccurate; if they were perfectly accurate, they would be complex. As such, the explanation must be wrong sometimes and is therefore not entirely trustworthy or transparent (Rudin, 2019).

3 Medicare data

Next, I discuss the data used. The analysis uses Medicare fee-for-service claims data, which include diagnoses and the amount paid for each service. Using these data, I closely, though not identically, follow the sample selection and variable creation procedures used in Medicare Advantage risk adjustment models. The predictor variables include demographics and health conditions in 2018. The outcome variable is annualized healthcare spending in 2019.

I restrict all models to standard Medicare risk adjustment variables so that they use variation already deemed acceptable by Medicare. As such, the new models differ from standard models primarily in their functional form, not in the variation they can use.

Appendix A provides more details on the data, sample selection, and variable construction.

The sample is split into a training set (80%), a validation set (10%), and a test set (10%). Models are fit in the training set, and model performance is currently evaluated on the validation set. The test set remains untouched and is available for future use.

The sample includes 4,002,909 individuals, of which 3,202,327 are included in the training sample. Appendix Tables D.2 and D.2 show summary statistics for the training and validation samples. The average patient has 2.60 (SD = 3.56) payment-relevant health conditions. The average annualized spending is $13,449.85 (SD = $35,441.55).

4 Model fitting and evaluation

I first fit risk adjustment models using standard and machine learning models. Then, for each model, I measure the model complexity using the number of coefficients and evaluate model accuracy using out-of-sample MAE. Last, I evaluate the trade-off between accuracy and complexity by estimating the marginal reduction in error per additional model coefficient.

4.1 Model specifications

Broadly, I fit three types of models: standard Medicare models, alternative linear models, and tree-based models.

Standard models: First, I refit standard Medicare models in my sample. This approach allows me to make a direct comparison between standard and new models that is uncontaminated by any potential differences in the underlying data used in model fitting. The fitted models include the Demographic model and the HCC model.

Alternative linear models: I include a selection of parametric models using OLS and lasso, designed to incorporate health condition interactions and reduce overfitting. To account for comorbidities, I fit an OLS model which interacts health conditions with counts

of other comorbidities. I refer to this as the "HCCxCount" model. I include both the HCC and HCCxCount variables in Lasso regression as well. Lasso regression is a standard model for reducing overfitting. It sets some coefficients to zero if they provide insufficient explanatory power while biasing the remaining coefficients toward zero (Hastie, Tibshirani, and Friedman, 2009).

Non-parametric (tree-based) machine learning models: I also fit non-parametric tree-based models—specifically, regression trees, random forests, and gradient-boosted trees. The single regression tree model is the simplest type of tree model. It splits the data into groups based on column values (e.g., males with heart failure and without diabetes) and generates a prediction for each group. Random forests fit multiple trees, each on random samples of the data and columns, and then average the predictions across trees. Random forests have performed well in risk adjustment models in commercial claims data (Rose, 2016). Gradient-boosted trees fit regression trees sequentially, with each tree fit on the residuals of the previous tree (Hastie, Tibshirani, and Friedman, 2009). Gradient-boosted trees have been used successfully to predict heart attacks, bail violations, and missed diagnoses (Mullainathan and Obermeyer, 2021; Kleinberg et al., 2018; Chan, Gentzkow, and Yu, 2022).

These models allow for flexible functional forms and variable interactions, allowing them to capture complex, non-linear interactions. Often, they yield higher-quality predictions than linear models (Hastie, Tibshirani, and Friedman, 2009; Rose, 2016). I train two versions of each model, one that minimizes MSE and another that minimizes MAE. Appendix B provides more details about model fitting and tuning.

4.2 Model accuracy

The primary accuracy metric I use is MAE, defined as the prediction's average distance from the realized value ($\sum_i^n \frac{|y_i - \hat{y}_i|}{n}$). A key advantage of this measure is its meaningful units: a 10-unit decrease in MAE indicates that predictions are, on average, $10 more accurate per

person. Additionally, MAE treats all errors equally, unlike metrics based on squared error. However, one disadvantage is that MAE differs from MSE, the metric that most linear models are trained on. To address this, in Appendix D I replicate the main results with MSE as the accuracy metric. In addition, for the main analysis, I assume that improved accuracy serves as a sufficient proxy for reduced selection incentives. The robustness of this assumption is evaluated in the results section.

To generate confidence intervals for model accuracy, I use a bootstrapping approach. Specifically, I generate 100 samples of the validation dataset, drawn with replacement. I then calculate the metrics of interest for each sample and use the 2.5% and 97.5% percentile values as 95% confidence interval bounds. This "quasi-Monte Carlo" approach, adapted from Park and Basu (2018), provides confidence intervals while preserving computational feasibility.

4.3 Model complexity

I measure model complexity as described previously. For linear models, model complexity is measured by the L0 norm, or the number of non-zero coefficients. For non-parametric, tree-based models, model complexity is measured as the number of unique and feasible decision paths. Equivalently, this measure is the number of combinations of variable values that lead to distinct predictions. For single trees, I measure this exactly as the number of "leaves," or terminal nodes, on the tree. For random forests and gradient-boosted trees, I estimate this as the number of unique predictions in the training data due to computational limitations. This estimate provides a lower bound to the model's complexity.

4.4 Marginal value of additional complexity

How should policymakers decide how to trade off complexity versus accuracy? For this analysis, I assume model complexity and accuracy are two of the many factors that policymakers value and that they have a constant marginal utility of both. I then consider relative

preferences between the two, holding all else equal.

The returns to complexity can be characterized as the marginal increase in accuracy per marginal increase in complexity. Policymakers will accept (or at least seriously consider) new models if the returns to complexity, in terms of error reduction, are sufficiently large.

Medicare's risk adjustment approach has evolved from the simpler Demographic model to the more accurate but complex HCC model. I use this transition to estimate bounds on preferences for complexity relative to accuracy. I estimate the marginal reduction in error per marginal coefficient from this change. I interpret this value as a revealed preference measure of error reduction per additional coefficient that policymakers are willing to accept.

Next, I estimate the error reduction per additional coefficient for new models. Focusing on models on the Pareto frontier, I compute the marginal reduction in error per marginal increase in coefficients for each model. I then compare this value with the value from past model changes. For robustness, I also consider alternative functional forms of preferences over complexity.

5 Results

In this section, I first estimate the accuracy and complexity of each model. One model significantly reduces error relative to the current Medicare model but it also substantially increases the number of coefficients. The reduction in error per additional coefficient is much less than that of past changes to Medicare models, suggesting that policymakers would likely not find this model preferable to the status quo. The results are robust to alternative accuracy measures, alternative preference assumptions, and upcoding. They are limited to the extent that accuracy is a sufficient proxy for selection incentives.

5.1 Model accuracy and complexity

Figure 1 displays the prediction accuracy for different models, and Figure 2 shows their varying levels of complexity. Figure 3 outlines the Pareto frontier of accuracy and complexity.

Among models on the Pareto frontier, complexity increases with accuracy. A prediction of the mean has a single parameter and therefore a complexity of 1. The average MAE is $15,792 (CI:15,683, 15,907). The Demographic model increases complexity to 13 coefficients and reduces MAE by $-540.20 (CI:$-553.70, $-525.90), or -3.4%. The HCC model adds more complexity, raising it to 113 coefficients. The additional complexity earns a larger decrease in MAE, reducing it by $-2,337 (CI:$-2,370, $-2,307), or -15%, relative to the mean.

The HCCxCount model achieves modest improvements in performance with modest increases in complexity. It increases complexity to 184. It also reduces MAE, but this reduction is not significantly different of that from HCC model ($-0.87; CI:$-7.24, $4.63).

The gradient-boosted tree trained on MAE reduces error the most, by $-3,690 (CI:$-3,714, $-3,667), or -23%, relative to the mean. This is a difference of $-1,352 (CI:$-1,392, $-1,316) relative to the HCC model. The change from the HCC model to the gradient-boosted tree is almost as large (roughly three-fourths the size) as the change from the Demographic to the HCC model, suggesting it is economically significant. However, this error reduction comes with an enormous degree of complexity: 187,389, or 165,831% that of the HCC model.

5.2 Marginal value of complexity

Is this increase in accuracy worth the large increase in complexity? To assess, I assume constant marginal utility for both accuracy and complexity. I use past changes in risk adjustment to infer bounds on relative preferences between the two. I find that policymakers would need to be willing to accept model changes that are considerably less efficient at reducing error than they have in the past for new models to be acceptable.

I first estimate a bound on acceptable changes. The HCC model reduces MAE by $17.97 per additional coefficient relative to the Demographic model. The transition from HCC to

Demographic model was adopted, which suggests that policymakers are willing to accept at least this rate of error reduction.

Figure 4 displays the marginal reduction in error per marginal increase in coefficients for Pareto models. The reduction in error falls with each additional increase in complexity.

The gradient-boosted tree (MAE) improves accuracy substantially, but inefficiently. Due to its complexity, it reduces error by a comparatively tiny amount per coefficient: \$0.00722 , relative to the next best model (and similarly, \$0.0072 relative to the HCC model). This is only 0.04% of reduction in error per additional coefficient from the transition to the HCC model. For this model to be acceptable, policymakers would have to be willing to accept a tiny fraction of the error reduction per coefficient than they have for past model changes.

5.3 Robustness

Next, I evaluate the robustness of these results with respect to alternative preference assumptions, measures of accuracy, and selection incentives. I find that they are largely robust to alternative assumptions that increase policymakers' tolerance of marginal complexity. I also find that they are robust to alternative measures of accuracy, but that accuracy is an imperfect proxy for selection incentives. Finally, I find that the improved accuracy is likely robust to upcoding.

5.3.1 Alternative preference assumptions

I consider alternative preference assumptions which progressively relax policymaker distaste for complexity. First, I assume policymakers care about the relative percentage increase in complexity. Then the return to complexity for the gradient-boosted tree is only 0.57% of the return from the switch to the HCC model. Second, I assume policymakers care about log complexity. Then the return to complexity is 23.5% of the return from the switch to the HCC model. Even with mild distaste for complexity, policymakers would need to accept much lower returns to complexity than before.

5.3.2 Alternative measures of accuracy

I next assess how the results change when using MSE to measure accuracy since MAE and MSE weight error differently. Appendix Figures D.1, D.2, and D.3 present the main analyses using MSE instead of MAE. The results are extremely similar. Again, gradient-boosted trees (this time trained on MSE) provide the largest reduction in MSE. The returns to complexity are remain low; policymakers would need to accept and error reduction per coefficient of 14.03, which is considerably less than past accepted changes of 1.403e+06.

5.3.3 Accuracy and selection incentives

Next, I assess the extent to which greater accuracy is a sufficient proxy for reduced selection incentives. Appendix Figure D.4 shows the performance of different models on a range of selection incentive measures. The measures are calculated overall and for patient subgroups thought to be subject to strong selection incentives, such as those with multiple chronic conditions (Zink and Rose, 2020; MedPAC, 2021)[2].

The extent to which accuracy proxies for selection incentives depends on the type of selection. For example, one selection incentive measure is the scope for selection conditional on risk score, as measured by the sum of positive residuals, which is highlighted by Brown et al. (2014). The gradient-boosted tree (MAE) substantially reduces the scope for this type of selection overall. It also greatly reduces differences across subgroups, reducing the incentive to select certain subgroups of patients. However, other metrics provide a different result. One such metric is tail risk, the probability that a patient's costs substantially exceed predictions, which is highlighted by Park and Basu (2018). I find that the gradient-boosted tree (MAE) increases tail risk overall and increases differences across subgroups, creating larger incentives to avoid certain subgroups. As such, the main results are limited to the extent that accuracy proxies for the dominant types of selection, which the literature is inconclusive on.

[2]Appendix C provides more details on the methods.

5.3.4 Robustness to Upcoding

One key limitation of the main analysis is that it considers model accuracy without considering strategic diagnosis coding. However, strategic diagnosis coding is prevalent in Medicare Advantage (Geruso and Layton, 2020). This limitation stems from using the Medicare claims data as they do not contain strategic diagnosis coding.

To understand how upcoding might affect model accuracy, I estimate the increase in predicted spending when adding a specific diagnosis to a patient's record, which also captures the returns from such upcoding. I focus on a subset of diagnoses listed in industry promotional material as the "biggest HCC coding opportunities" (RCX Rules, 2023). Figure 5 shows the distributions of predicted spending increases for both the HCC model and gradient-boosted tree trained on MAE. Appendix Figure D.5 replicates this with the gradient-boosted tree trained on MSE.

I find that the HCC model consistently shows higher increases in predicted spending. For the six diagnoses, upcoding in the HCC model raises spending by several thousand dollars more than in the gradient-boosted tree. This observation suggests that the gradient-boosted trees generate more stable, and therefore more accurate, predictions in the presence of upcoding and lower incentives to upcode.

6 Discussion

Gradient-boosted trees improve accuracy relative to standard risk adjustment models, consistent with results from other settings and features of gradient-boosted trees. However, they increase complexity to a degree that they are unlikely to be acceptable to policymakers. Key limitations of this work are that it does not test all potential model functional forms and uses accuracy as a proxy for varied selection incentives.

Why do the gradient-boosted trees offer such large increases in accuracy and complexity, and to what extent is this surprising? Regarding accuracy, gradient-boosted trees have

consistently proven to be the best at structured machine learning problems. For example, at Kaggle, an online platform where people compete to solve machine learning problems, gradient-boosted trees were found to win the most competitions for supervised learning problems across a range of domains (Harasymiv, 2015; Nielsen, 2016). Regarding complexity, gradient-boosted trees impose minimal functional form assumptions, which allows them to capture non-linear relationships but also requires many more parameters. In addition, they are ensembles of multiple trees, which greatly increases parameters relative to using a single tree or linear model.

Another reason why the gradient-boosted trees perform well is that they are fit to optimize MAE, while the linear models are optimizing MSE. While this could seem an unfair comparison, it highlights a key strength of this type of model. Standard linear models are largely tied to the objective of minimizing MSE. In contrast, gradient-boosted trees can target any custom objective that is a function of y and $\hat{y}$, leading them to perform better at that objective. This is a useful feature given recent literature suggesting alternative objective functions for risk adjustment (Layton, McGuire, and Van Kleef, 2018; Zink and Rose, 2020).

The results show that for gradient-boosted trees to be acceptable, policymakers would need to accept substantially lower returns to complexity than they have in the past. This finding is robust to alternative preference assumptions that reduce the marginal distaste for complexity and to different accuracy metrics. Thus, more accurate machine learning models can be reasonably rejected as improvements over current models, assuming policymakers' dislike of complexity has not fallen dramatically over time. Policymakers would need to be over a hundred times more tolerant of complexity than they have currently revealed themselves to be for these models to be acceptable.

One limitation of this study is that I have not tested every possible functional form that could use these variables. While other models (or versions of these models) may exist that provide similar accuracy for much less complexity, they would need to provide the observed increase in accuracy with less than 1/100th of the additional complexity of the models I

examine. This degree of improvement seems unlikely to result from functional form changes alone. Therefore, standard machine learning methods are likely insufficient to substantially improve risk adjustment accuracy without increasing complexity to unacceptable levels.

Another limitation is that accuracy is an imperfect proxy for varied types of selection incentives. In line with past literature, I find that the optimal model depends on the relative importance of different types of selection and the data moment used to proxy it (Park and Basu, 2018; Zink and Rose, 2020). While this study is not the first to observe the limitations of minimizing prediction error as an approach to risk adjustment (Layton, McGuire, and Van Kleef, 2018; Zink and Rose, 2020; Geruso and Layton, 2017), this method remains the conventional approach (MedPAC, 2021). Geruso and Layton (2017) review a number of these challenges and conclude that despite its limitations, conventional risk adjustment remains among the best tools available to address selection, as evidenced by its wide adoption.

7 Conclusion

This paper examines the trade-off between accuracy and interpretability in risk adjustment models for Medicare Advantage. I find that although interpretability is a documented criterion, it has been previously informal and unquantified in the context of payment policy. To address this, I introduce a concrete measure of model complexity (or non-interpretability). By quantifying interpretability in this context, I formalize an important auxiliary criterion, enabling the formal objective to be more fully specified. Using the same data and variables employed in Medicare Advantage risk adjustment, I assess both traditional and machine learning models. My analysis reveals that machine learning models can significantly enhance prediction accuracy and improve robustness to upcoding but introduce unprecedented levels of complexity. These models provide very small increases in accuracy relative to their increase in complexity. As such, they would require policymakers to accept substantially smaller reductions in error per additional coefficient than they have in the past. Most likely,

policymakers would not consider these models improvements over the status quo. As such, when accounting for auxiliary criteria in policy contexts, like interpretability, the optimal choice of models is likely to change meaningfully. Consequently, future research should explore how to use the advances offered by machine learning in ways that align with policy constraints in this critical domain.

8 Appendix

References

Berry, Christopher R. 2021. "Reassessing the property tax." *Available at SSRN 3800536*.

Bertsimas, Dimitris, Angela King, and Rahul Mazumder. 2016. "Best Subset Selection Via a Modern Optimization Lens." *The Annals of Statistics*, 813–852.

Brown, Jason, Mark Duggan, Ilyana Kuziemko, and William Woolston. 2014. "How Does Risk Selection Respond to Risk Adjustment? New Evidence from the Medicare Advantage Program." *American Economic Review*, 104(10): 3335–3364.

CBO. 2022. "The Prices That Commercial Health Insurers and Medicare Pay for Hospitals' and Physicians' Services." Congressional Budget Office 57422.

Chan, David C., Matthew Gentzkow, and Chuan Yu. 2022. "Selection with Variation in Diagnostic Skill: Evidence from Radiologists." *The Quarterly Journal of Economics*, 137(2): 729–783.

Chen, Tianqi, and Carlos Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *KDD '16*, 785–794. New York, NY, USA:ACM.

CMS. 2023. "CMS-HCC Software V2422.86.P2." *https://www.cms.gov/files/zip/2023-initial-icd-10-mappings.zip*, accessed April 28, 2023.

Cubanski, Juliette, and Tricia Neuman. 2023. "The Facts on Medicare Spending and Financing." *Henry J. Kaiser Family Foundation, San Francisco*.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." *arXiv preprint arXiv:1702.08608*.

Ellis, Randall P., and Thomas G. McGuire. 2007. "Predictability and Predictiveness in Health Care Spending." *Journal of Health Economics*, 26(1): 25–48.

Geruso, Michael, and Timothy J. Layton. 2017. "Selection in Health Insurance Markets and Its Policy Remedies." *Journal of Economic Perspectives*, 31(4): 23–50.

Geruso, Michael, and Timothy Layton. 2020. "Upcoding: Evidence from Medicare on Squishy Risk Adjustment." *Journal of Political Economy*, 128(3): 984–1026.

Gottlieb, Daniel J., Weiping Zhou, Yunjie Song, Kathryn Gilman Andrews, Jonathan S. Skinner, and Jason M. Sutherland. 2010. "Prices Don't Drive Regional Medicare Spending Variations." *Health Affairs*, 29(3): 537–543.

Harasymiv, Vasyl. 2015. "Lessons from 2 Million Machine Learning Models on Kaggle."

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2, Springer.

Hazimeh, Hussein, and Rahul Mazumder. 2020. "Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms." *Operations Research*, 68(5): 1517–1537.

Henriquez, Josefa, Richard C. van Kleef, Andrew Matthews, Thomas McGuire, and Francesco Paolucci. 2023. "Combining Risk Adjustment with Risk Sharing in Health Plan Payment Systems: Private Health Insurance in Australia." National Bureau of Economic Research.

Irvin, Jeremy A., Andrew A. Kondrich, Michael Ko, Pranav Rajpurkar, Behzad Haghgoo, Bruce E. Landon, Robert L. Phillips, Stephen Petterson, Andrew Y. Ng, and Sanjay Basu. 2020. "Incorporating Machine Learning and Social Determinants of Health Indicators into Prospective Risk Adjustment for Health Plan Payments." *BMC Public Health*, 20: 1–10.

James, Melissa, Michael Stearns, and Kimberly Rykaczewski. 2023. "A First Look at the 2024 CMS Advance Notice." *https://www.wolterskluwer.com/en/expert-insights/a-first-look-at-the-2024-cms-advance-notice*, accessed June 7, 2023.

Kan, Hong J., Hadi Kharrazi, Hsien-Yen Chang, Dave Bodycombe, Klaus Lemke, and Jonathan P. Weiner. 2019. "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*, 14(3): e0213258.

Kautter, John, Gregory C Pope, and Patricia Keenan. 2014. "Affordable Care Act Risk Adjustment: Overview, Context, and Challenges." *Medicare & Medicaid Research Review*, 4(3).

KFF. 2023. "Total Medicaid MCO Spending." Accessed on May 25, 2023.

Kleinberg, Jon, and Sendhil Mullainathan. 2019. "Simplicity Creates Inequity: Implications for Fairness, Stereotypes, and Interpretability." 807–808.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics*, 133(1): 237–293.

Layton, Timothy J., Thomas G. McGuire, and Richard C. Van Kleef. 2018. "Deriving Risk Adjustment Payment Weights to Maximize Efficiency of Health Insurance Markets." *Journal of Health Economics*, 61: 93–110.

McGuire, Thomas G., Anna L. Zink, and Sherri Rose. 2020. "Simplifying and Improving the Performance of Risk Adjustment Systems." National Bureau of Economic Research Working Paper 26736.

McGuire, Thomas G., Joseph P. Newhouse, and Anna D. Sinaiko. 2011. "An Economic History of Medicare Part C." *The Milbank Quarterly*, 89(2): 289–332.

Medicare Trustees. 2022. "2022 Annual Report of the Boards of Trustees of the Federal Hospital Insurance and Federal Supplementary Medical Insurance Trust Funds."

MedPAC. 2014. "Report to the Congress: Medicare and the Health Care Delivery System." Medicare Payment Advisory Commission.

MedPAC. 2021. "Report to the Congress: Risk Adjustment in Medicare Advantage." Medicare Payment Advisory Commission.

Mullainathan, Sendhil, and Ziad Obermeyer. 2021. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *The Quarterly Journal of Economics*, 137(2): 679–727.

Newhouse, Joseph P., Mary Price, John Hsu, J. Michael McWilliams, and Thomas G. McGuire. 2015. "How Much Favorable Selection Is Left in Medicare Advantage?" *American Journal of Health Economics*.

Nielsen, Didrik. 2016. "Tree Boosting with XGBoost—Why Does XGBoost Win "Every" Machine Learning Competition?" Master's diss. NTNU.

Norregaard, Mr John. 2013. *Taxing immovable property revenue potential and implementation challenges.* International Monetary Fund.

Park, Sungchul, and Anirban Basu. 2018. "Alternative Evaluation Metrics for Risk Adjustment Methods." *Health Economics*.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, 12: 2825–2830.

Pope, Gregory C., John Kautter, Randall P. Ellis, Arlene S. Ash, John Z. Ayanian, Lisa I. Iezzoni, Melvin J. Ingber, Jesse M. Levy, and John Robst. 2004. "Risk Adjustment of Medicare Capitation Payments Using the CMS-HCC Model." *Health Care Financing Review*, 25(4): 119.

RCX Rules. 2023. "An Overview of Medicare." Accessed September 25, 2023.

ResDAC. 2023. "30 CCW Chronic Conditions Algorithms: MBSF_CHRONIC_{YYYY}." Centers for Medicaid and Medicare.

Rose, Sherri. 2016. "A Machine Learning Framework for Plan Payment Risk Adjustment." *Health Services Research*, 51(6): 2358–2374.

Rose, Sherri, Savannah L. Bergquist, and Timothy J. Layton. 2017. "Computational Health Economics for Identification of Unprofitable Health Care Enrollees." *Biostatistics*, 18(4): 682–694.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence*, 1(5): 206–215.

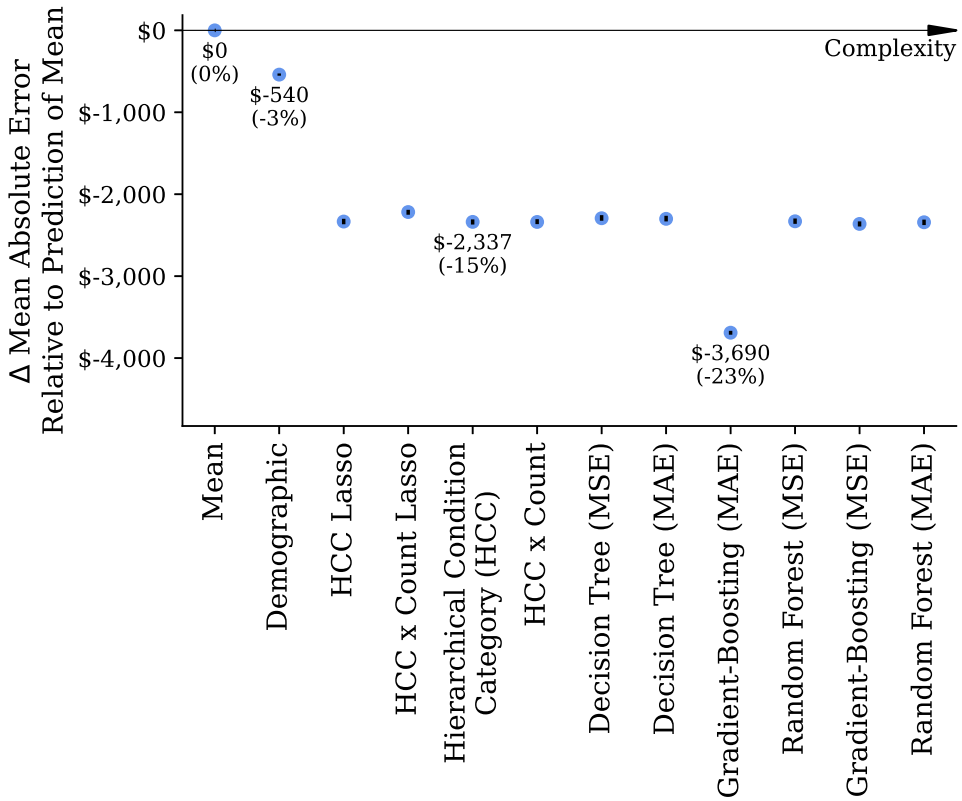
Ryan, Andrew M., Zoey Chopra, David J. Meyers, Erin C. Fuse Brown, Roslyn C. Murray, and Travis C. Williams. 2023. "Favorable Selection in Medicare Advantage Is Linked to Inflated Benchmarks and Billions in Overpayments to Plans: Study Examines Medicare Advantage Favorable Selection, Benchmarks, and Payments to Plans." *Health Affairs*, 42(9): 1190–1197.

Statista Research Department. 2022. "U.S. Gross Domestic Product: Forecast 2021-2032." Accessed: 2023-05-25.

Zhu, Jane M., Mark Katz Meiselbach, Coleman Drake, and Daniel Polsky. 2023. "Psychiatrist Networks in Medicare Advantage Plans Are Substantially Narrower Than in Medicaid and ACA Markets." *Health Affairs*, 42(7): 909–918. PMID: 37406238.

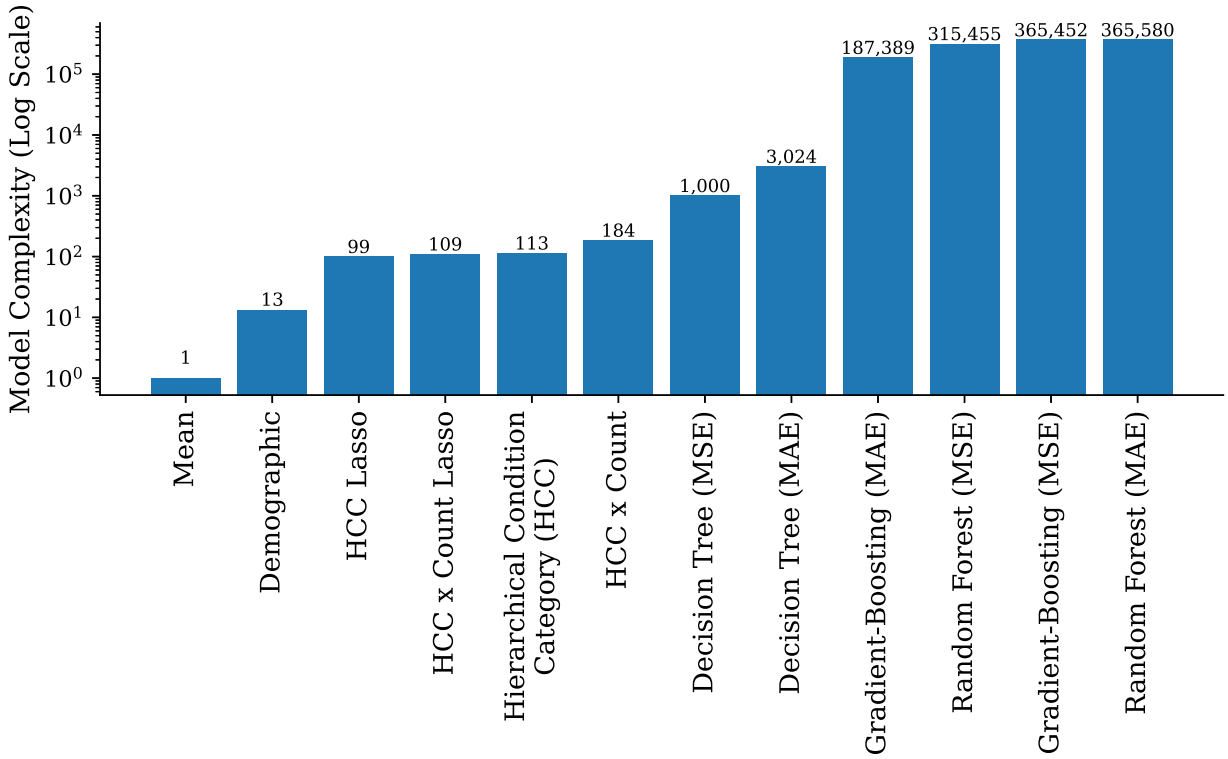
Zink, Anna, and Sherri Rose. 2020. "Fair Regression for Health Care Spending." *Biometrics*, 76(3): 973–982.

Figure 1: Difference in MAE of Model Predictions Relative to Predicting the Mean



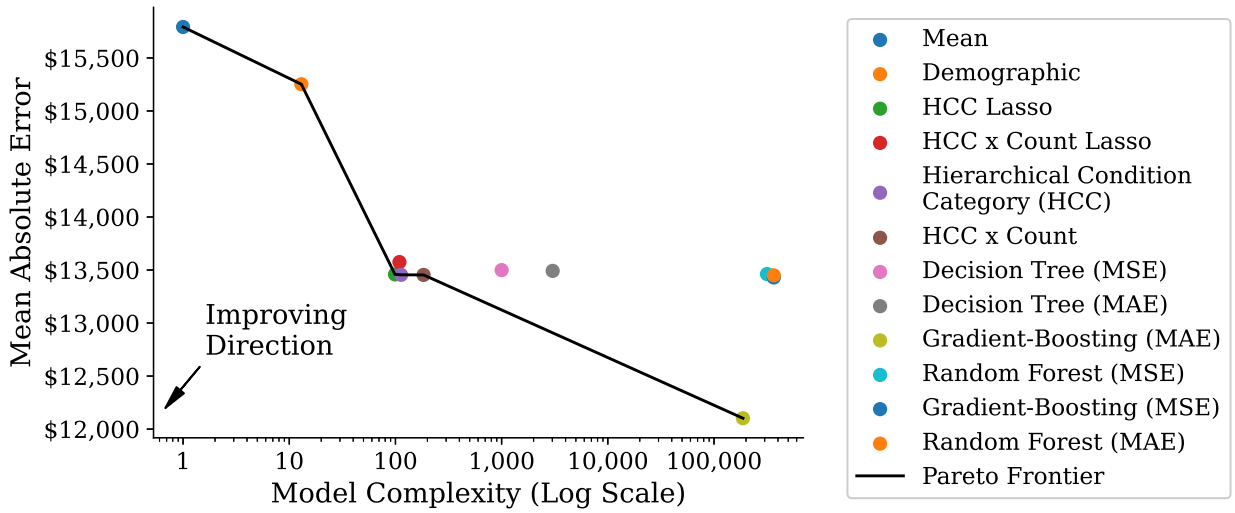
Notes: This graph shows the change in MAE for model predictions relative to predicting the mean. The x-axis shows the names of various models, ordered by increasing complexity. The "Hierarchical Condition Category" model is the current Medicare risk adjustment model. Models with "MAE" or "MSE" after their name indicate the objective function that the model was trained on, mean absolute error or mean squared error. The y-axis shows the reduction in the out-of-sample MAE, relative to always predicting the mean. MAE is calculated out of sample in 10% of the available data. The point estimate of MAE for each model is represented by the round marker. Standard errors are calculated with bootstrapped samples in the out-of-sample data. Ninety-five percent confidence intervals are shown as black bars; note that they are narrower than the height of the round markers. The values in parentheses are the percentage reductions in MAE relative to always predicting the mean.

Figure 2: Model Complexity



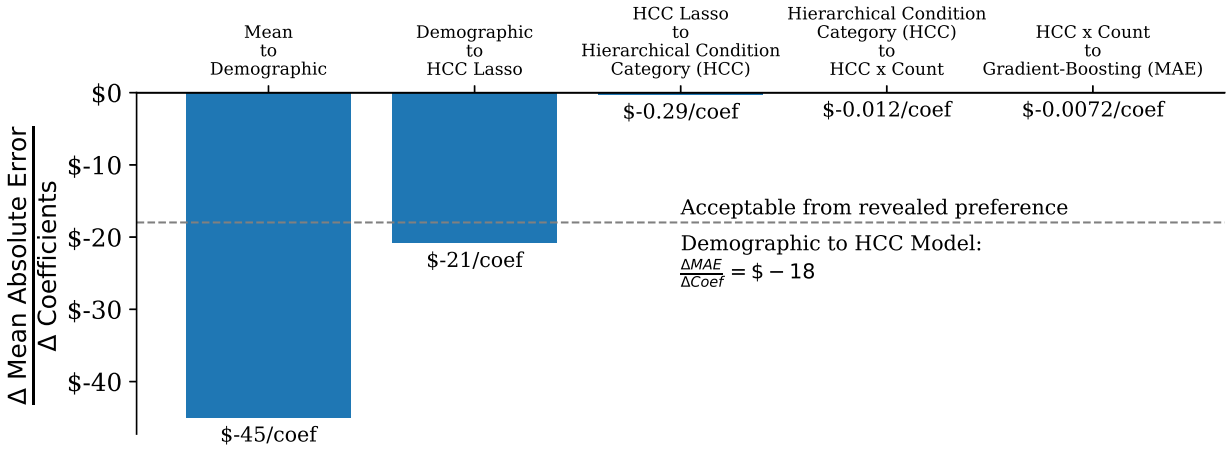
Notes: This graph shows the complexity of selected risk adjustment models fit in the data. The x-axis shows the names of various models. The "Hierarchical Condition Category (HCC)" model is the current Medicare risk adjustment model. Models with "MAE" or "MSE" after their name indicate the objective function that the model was trained on, mean absolute error or mean squared error. The y-axis shows the "complexity," or non-interpretability, of a model. Model complexity is calculated as the number of coefficients required to characterize the range of outputs. 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.

Figure 3: Pareto Frontier of Accuracy (MAE) and Complexity (Number of Coefficients)



Notes: This graph shows a scatterplot of the accuracy and complexity of different risk adjustment models. The x-axis, in log scale, shows model complexity, measured as the number of coefficients. The y-axis shows model accuracy as measured by MAE out of sample. The black line shows the Pareto frontier of accuracy and complexity. The improving direction is toward the origin, or left and down, toward zero MAE and zero complexity.

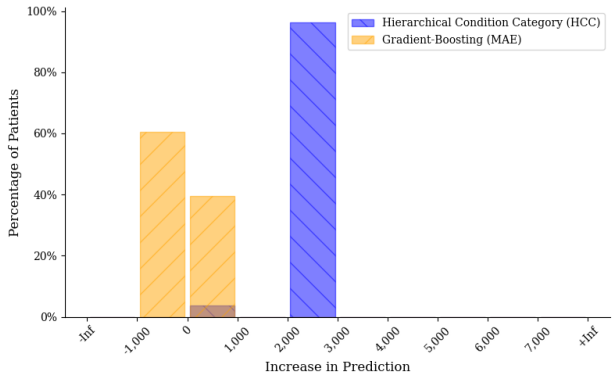
Figure 4: Marginal Change in MAE per Coefficient by Model for Subset of Pareto Models in Terms of Complexity and MAE



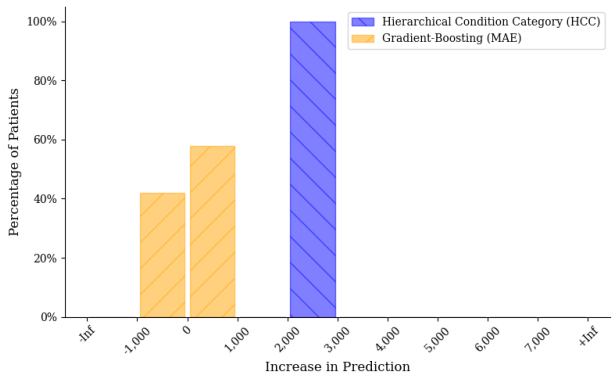
Notes: This figure restricts attention to models that are on the Pareto frontier of MAE and complexity. The x-axis lists model pairs in order of increasing complexity, and the y-axis shows the marginal decrease in MAE per additional model coefficient for each pair of models. The dotted horizontal gray line shows the marginal decrease in MAE per additional model coefficient for past risk adjustment model changes, specifically the change from the Demographic to the HCC model. These past changes were acceptable to policymakers. The more accurate, more complex models offer a much smaller decrease in MAE per additional coefficient. For these new models to be acceptable, policymakers would have to be willing to accept much smaller error reduction per coefficient than they have in the past.

Figure 5: Increase in Predicted Patient Cost from Upcoding

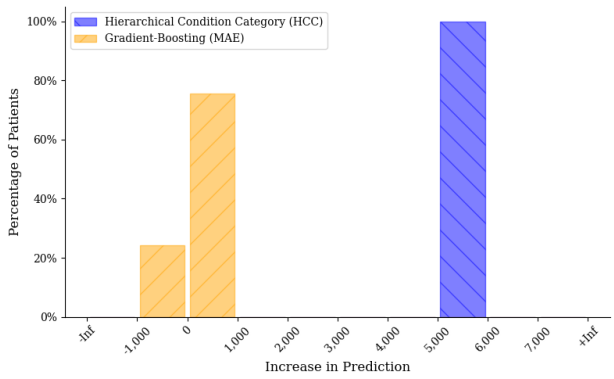
(a) Diabetes with Chronic Conditions (HCC 19)



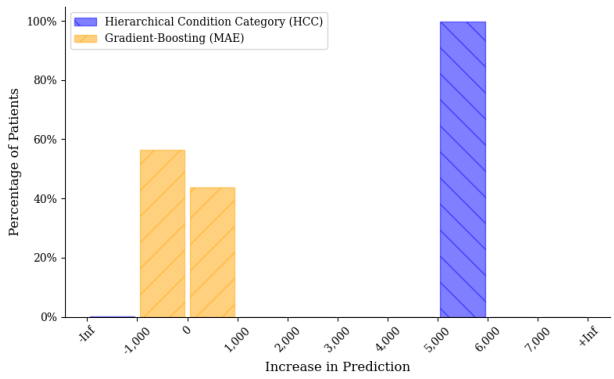
(b) Morbid Obesity (HCC 22)



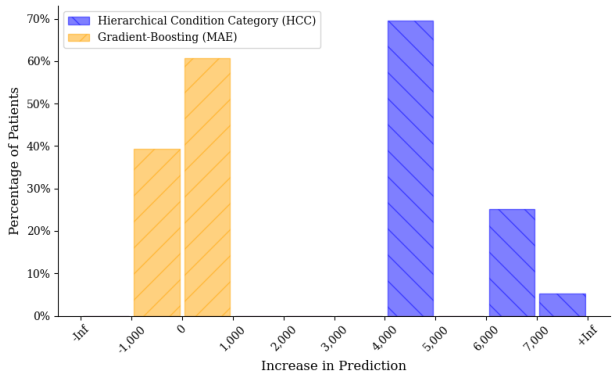
(c) Rheumatoid Arthritis (HCC 40)



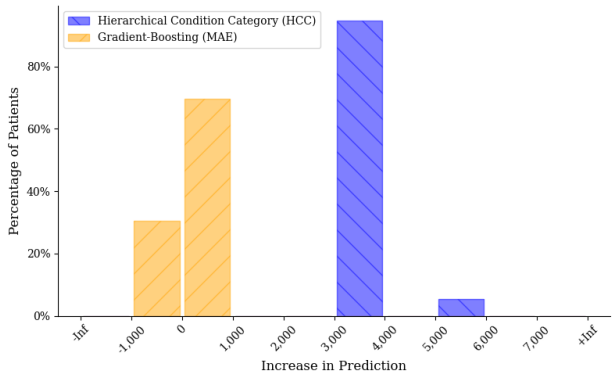
(d) Coagulation Defects (HCC 48)



(e) Congestive Heart Failure (HCC 85)



(f) Specified Heart Arrhythmias (HCC 96)



Notes: Panel (a) shows the distribution of the change in predicted patient costs from adding diabetes with chronic complications (HCC 19) to patient records without HCC 19. 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. 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. Panels (b), (c), (d), (e), and (f) show the same analysis for adding morbid obesity (HCC 22), rheumatoid arthritis (HCC 40), coagulation defects (HCC 48), congestive heart failure (HCC 85), and heart arrhythmias (HCC 96), respectively.

9 Appendix

For Online Publication

Accuracy and Complexity in Payment Algorithms

Maya Lozinski

October 2023

A Sample and variable construction

Data overview. The Medicare data contain records of healthcare usage and diagnoses. I use diagnoses from the Carrier (physician services), Home Health, Outpatient (outpatient facility fees), and MedPAR (inpatient facility fees) claims files. I calculate spending from these same files.

Sample selection. The sample selection is as follows. Individuals must have been enrolled in both Medicare Part A and B for all of 2018 and at least one month of 2019. They are included if they are 65 years or older and qualified for Medicare by age, and they are excluded if they have end-stage renal disease or enrolled in Medicaid or Medicare Advantage at any time between 2018 and 2019. Individuals are also excluded if their gender is unknown or their state of residence is not a US state (e.g., a territory). Medicare also restricts its sample to those for whom Medicare is a primary payer and subsets the sample based on whether a patient is institutionalized long term. I cannot observe these variables, but I assume the vast majority of patients have Medicare as a primary payer and are not institutionalized.

Spending calculations. To calculate total patient spending, I add up all spending by Medicare, the patient, and other sources across all files in 2019. Spending is annualized by the number of months the patient was enrolled in Medicare Part A and B. I do not perform any price adjustments. Prices in Medicare are administratively set and vary slightly across geography due to geographic adjustments and other rate-setting tools. However, this price variation is small compared to variation in private insurance prices, and it does not drive

meaningful variation in spending (CBO, 2022; Gottlieb et al., 2010).

Note that Medicare risk adjustment takes the additional step of dividing total spending by mean spending such that spending outcomes are a percentage of mean spending (e.g., 200% of the mean). I do not implement this step, which keeps outcome units in 2019 dollars and aids in interpretability. The difference does not otherwise affect the results.

Predictor variable structure for non-parametric models. The machine learning models use the same variables as the standard HCC model. However, certain variables are formatted differently for the non-parametric models (tree, random forest, gradient-boosted tree) than in the HCC linear model. Broadly, variables are formatted as single variables (e.g., age, sex) rather than as a series of saturated indicators with pre-specified interactions.

- **Age.** Age information is binned into the same groupings as used in the HCC model. However, age group is provided to the model as a single ordinal variable, containing an ordered group number, rather than as a series of indicator variables interacted with sex. This preserves the ordinal information in the age group variable.

- **Sex.** Sex is provided to the machine learning models as a single binary variable rather than as a series of indicator variables interacted with the age group. This allows sex to be interacted with any variable in the process of model fitting rather than restricting it to interactions with the age group.

- **HCC groupings.** Each HCC grouping (e.g., diabetes of any severity) is provided to the machine learning models as a single indicator variable rather than as a series of interaction variables between specified HCC groupings. This allows HCC groupings to be interacted with any other HCC grouping or variable rather than restricting the interaction to a specified subset of other HCC groupings.

B Model fitting and tuning

This section discusses the implementation of risk adjustment model fitting in greater detail to aid in replicability. The code is available upon request.

B.1 Standard Medicare Models

Demographic and HCC Models. The Demographic model uses five-year age bins interacted with sex indicators. This HCC model adds 86 hierarchical health condition indicators, 7 health condition interactions, and health condition count indicator variables for counts 4 through 10+. I use the 2023 V24 model version and fit it for community-dwelling, aged, non-disabled, non-Medicaid beneficiaries (CMS, 2023).

Validation of recalibrated Medicare models To verify the reliability of the Medicare models used in this analysis, I confirm that the recalibrated Medicare models have a very similar in-sample R^2 to the R^2 reported by Medicare. Medicare reports indicate an R^2 of 0.77% for the Demographic model, slightly lower than the R^2 of 1.61% in this sample. For the V24 HCC model, Medicare reports indicate an R^2 of 12.57, slightly higher than the R^2 of 11.03% in this sample (MedPAC, 2021). I attribute the small differences in R^2 to differences in sample year and construction and to the inherent variance in the R^2 estimates.

B.2 Alternative linear models:

I include a selection of parametric models using OLS and lasso, designed to incorporate health condition interactions and reduce overfitting.

Lasso. The first model employs lasso regression with standard HCC model coefficients. Lasso regression is a standard model for reducing overfitting and therefore improving out-of-sample performance. Like linear regression, it minimizes MSE but adds a constraint on the max value of the sum of the absolute values of coefficients. As a result, it sets some coefficients to zero if they provide insufficient explanatory power while biasing the remaining

coefficients toward zero (Hastie, Tibshirani, and Friedman, 2009).

Of note, the lasso implements an L1 penalized sparse regression (limits the sum of the absolute value of coefficients). The ideal regression here would implement an L0 constraint, which limits the number of non-zero coefficients. However, L0 regression is notoriously computationally inefficient and likely infeasible on a dataset of this size (Bertsimas, King, and Mazumder, 2016; Hazimeh and Mazumder, 2020).

Lasso regressions are fit using Scikit-Learn (Pedregosa et al., 2011). The optimal sparsity parameter for lasso is determined via cross-validation using the approach implemented in Pedregosa et al. (2011), "LassoCV."

"HCCxCount" model Another variant I introduce is "HCCxCount," with linear and lasso regression versions. This set of models is motivated by the observation that Medicare risk adjustment still consistently underpredicts costs for patients with multiple comorbidities and is known to inadequately account for comorbidity interactions (MedPAC, 2021). These models add a set of regression variables that multiply each HCC indicator with the total count of HCCs, allowing the effect of health conditions on costs to increase with a patient's overall comorbidity burden. They also include standard HCC, HCC interaction, HCC count, and demographic variables. I fit both a standard linear regression and a lasso regression with these variables.

B.3 Tree-Based Machine Learning Models

All machine learning models are fit using Scikit-Learn (Pedregosa et al., 2011) except for gradient-boosted trees, which are fit using XGBoost (Chen and Guestrin, 2016).

Regression tree and random forest. The hyperparameters are determined via a random search of hyperparameter values, and selected hyperparameters are those that generate models that perform best in threefold cross-validation. The models are trained to minimize either MSE or MAE, and they are evaluated in cross-validation accordingly. Once the best set of hyperparameters are chosen, the model is refit with these parameters on the full

training dataset.

Gradient-boosted regression tree. The hyperparameter tuning process is the same as for regression trees and random forest, with one additional step. Once the best set of hyperparameters are chosen, the model is refit with these parameters on the full training dataset with early stopping criteria. Early stopping prevents the model from fitting additional trees once additional trees stop improving out-of-sample fit.

C Selection incentives analysis

The main text's analysis focuses on the trade-off between complexity and accuracy. In that context, I assume that a lower MAE serves as a sufficient statistic for both improved accuracy and reduced selection incentives. Figure D.4 evaluates the validity of this assumption.

In addition to estimating MAE, I also calculate several common measures of selection incentives pulled from the literature, both overall and for select patient subgroups. This section describes the methods for this analysis and key results.

I first calculate MAE for different subgroups of patients where there is plausible means and motivation for selection. Evaluated subgroups include individuals with chronic mental health disorders, chronic substance use disorders, multiple chronic conditions, and no chronic conditions. The first three consistently cost more than predicted, while the last one consistently costs less (Zink and Rose, 2020; MedPAC, 2021).

To construct these groups, I use "chronic condition" variables taken from the Master Beneficiary Summary File Chronic Conditions and Other Chronic Conditions files. These variables are constructed using diagnoses from multiple years of claims data and prescription information (ResDAC, 2023). They differ from the standard HCC health condition variables, which use only diagnoses from one year of claims data. As such, these chronic condition variables contain additional information that likely predicts spending and may be available to insurance plans. However, because these variables are not included in risk adjustment

models, they reflect dimensions along which insurance companies have both the incentives and means to influence patient selection. These considerations make them important factors for assessing selection incentives.

Figure D.4a shows the MAE for these patient subgroups for different risk adjustment models. Overall, the MAE within each subgroup largely declines as model complexity increases, reaching its lowest with the gradient-boosted tree (MAE).

The first type of selection incentive I consider is selection conditional on predicted risk. Some patients may incur a lot of costs, yet their predicted costs could be even higher. These patients represent opportunities for cream-skimming, or positive selection. The sum of positive residuals captures the potential opportunity for positive selection (Brown et al., 2014). Figure D.4b shows the results by subgroup. Interestingly, the sum of positive residuals holds roughly constant or declines as model complexity increases. It drops dramatically with the gradient-boosted tree (MAE), both overall and by subgroup, and leads to much smaller differences across subgroups.

The second type of selection is selection to avoid tail risk. In a world where insurance companies still have some degree of risk aversion, they will want to avoid patients with a high risk of costing substantially more than predicted and be less concerned about small deviations from predicted costs. Figure D.4c shows the probability that a patient costs over twice as much as predicted, following Park and Basu (2018). Broadly, differences in tail risk across groups are narrow with more complex models. The one exception to this is the gradient-boosted tree (MAE), which raises tail risk particularly for high-cost groups and leads to larger group-level differences.

The last type of selection incentive I examine is selection by risk, e.g., by expected compensation. Some patient groups consistently incur costs that exceed their predicted values. This discrepancy is typically measured as the net compensation for a group, representing the expected loss or gain per patient ($\frac{\sum_i \hat{y}_i - \sum_i y_i}{N}$). Accordingly, I calculate net compensation for various subgroups under different risk adjustment models. Figure D.4d shows the

results. The HCC model successfully narrows differences across groups in net compensation relative to predicting the mean. The gradient-boosted tree (MAE) leads to negative net compensation overall and larger differences across subgroups.

D Additional exhibits

Table D.1: Training Sample Summary Statistics

Variable	Mean	StdDev
Age	75.62	7.50
Female	0.56	0.50
Count of HCCs	2.60	3.56
Annualized Spending	13,449.85	35,441.55
Mortality Rate	0.04	0.19

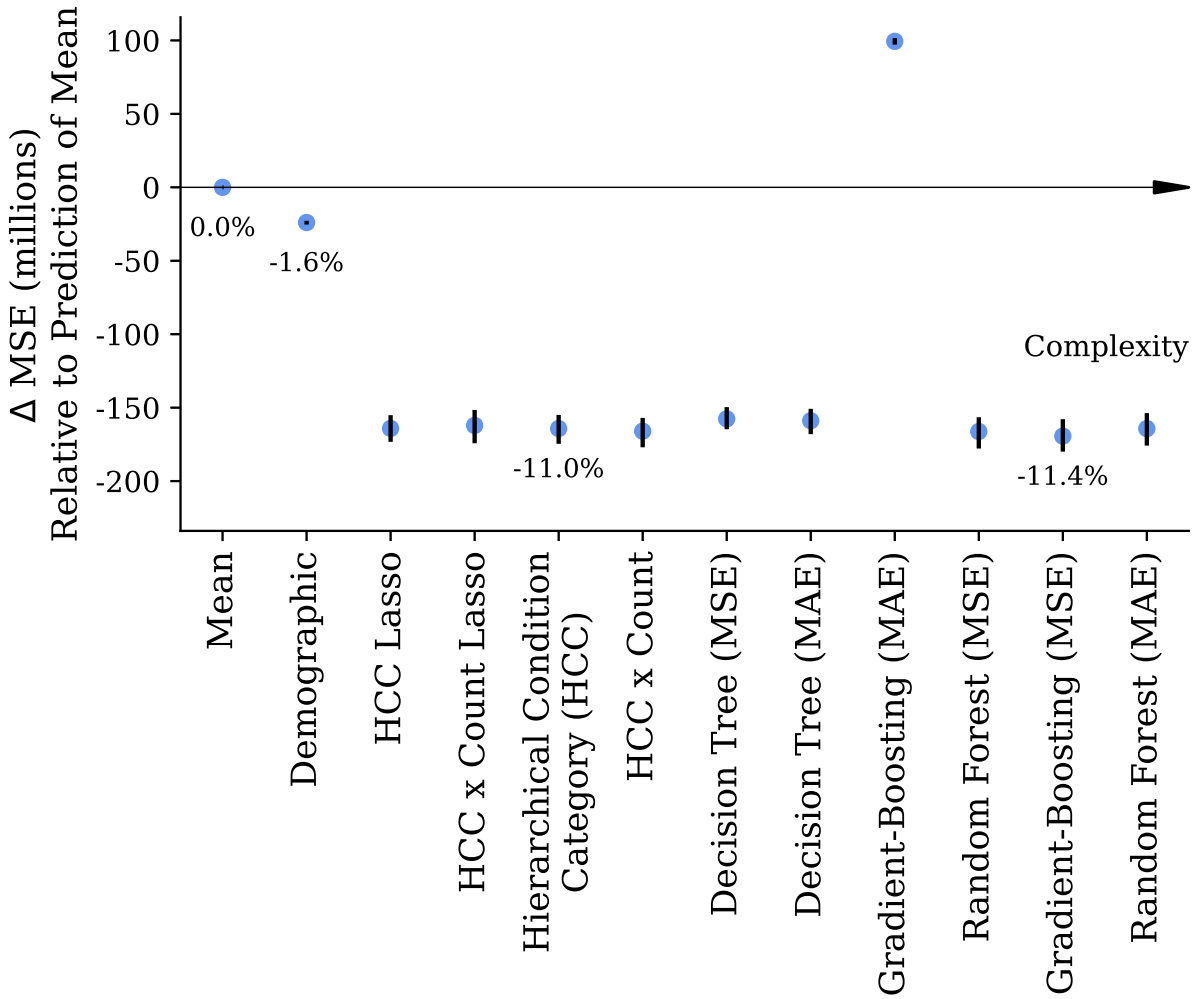
Notes: This table shows summary statistics for the data the models are trained on. The count of HCCs is the count of HCCs per person. These are determined based on diagnosis codes in claims in 2018, the "base year." Annualized spending and death rates are calculated in 2019, the "outcome year" for prospective payments.

Table D.2: Validation Sample Summary Statistics

Variable	Mean	StdDev
Age	75.61	7.50
Female	0.56	0.50
Count of HCCs	2.61	3.57
Annualized Spending	13,512.97	38,587.86
Mortality Rate	0.04	0.19

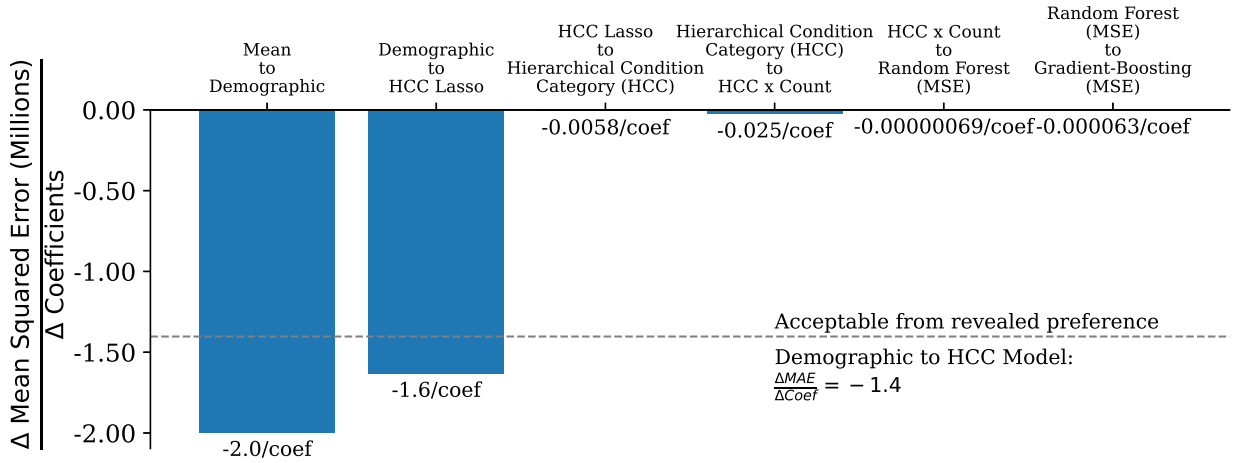
Notes: This table shows summary statistics for the validation data, i.e., the data where model accuracy is assessed out of sample. It is analogous to Table , which shows summary statistics for the training data. The count of HCCs is the count of HCCs per person and is determined based on diagnosis codes in claims in 2018, the "base year." Annualized spending and death rates are calculated in 2019, the "outcome year" for prospective payments.

Figure D.1: Difference in MSE of Model Predictions Relative to Predicting the Mean



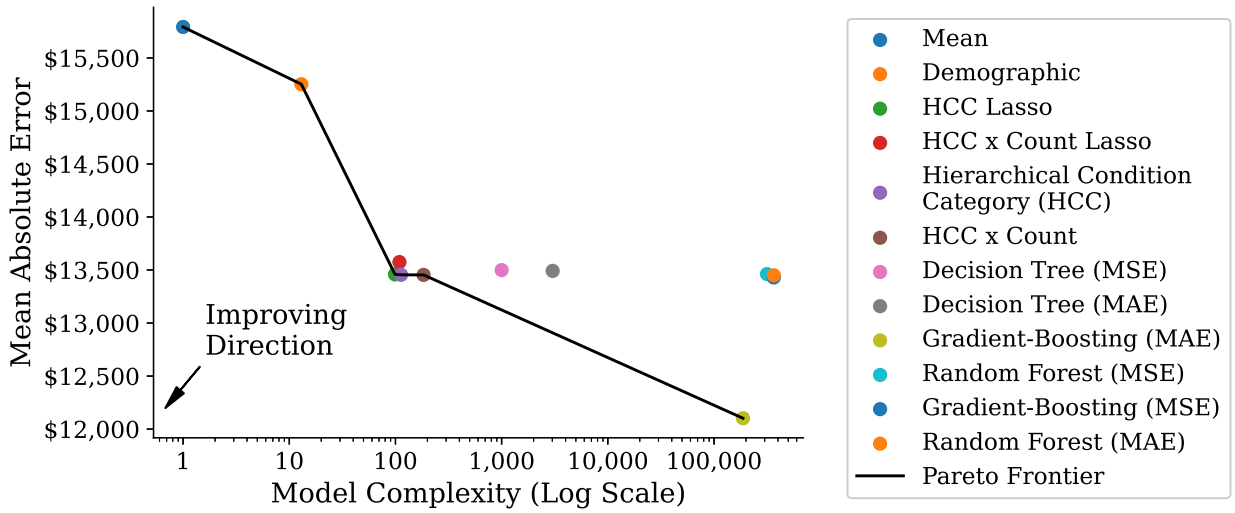
Notes: This graph shows the change in the MSE for model predictions relative to predicting the mean. It is analogous to Figure , but shows MSE rather than MAE. The x-axis shows the names of various models, ordered by increasing complexity. The "Hierarchical Condition Category" model is the current Medicare risk adjustment model. Models with "MAE" or "MSE" after their name indicate the objective function on which the model was trained, the mean absolute error, or the mean squared error. The MSE is calculated out of sample in 10% of the available data. The point estimate is the round marker, and standard errors are calculated with bootstrapped samples in the out-of-sample data. Ninety-five percent confidence intervals are shown as black bars; note that in some cases they are narrower than the height of the round markers. The values in parentheses are the percentage reductions in the MSE relative to always predicting the mean.

Figure D.2: Marginal Change in MSE per Coefficient by Model for Subset of Pareto Models in Terms of Complexity and MAE



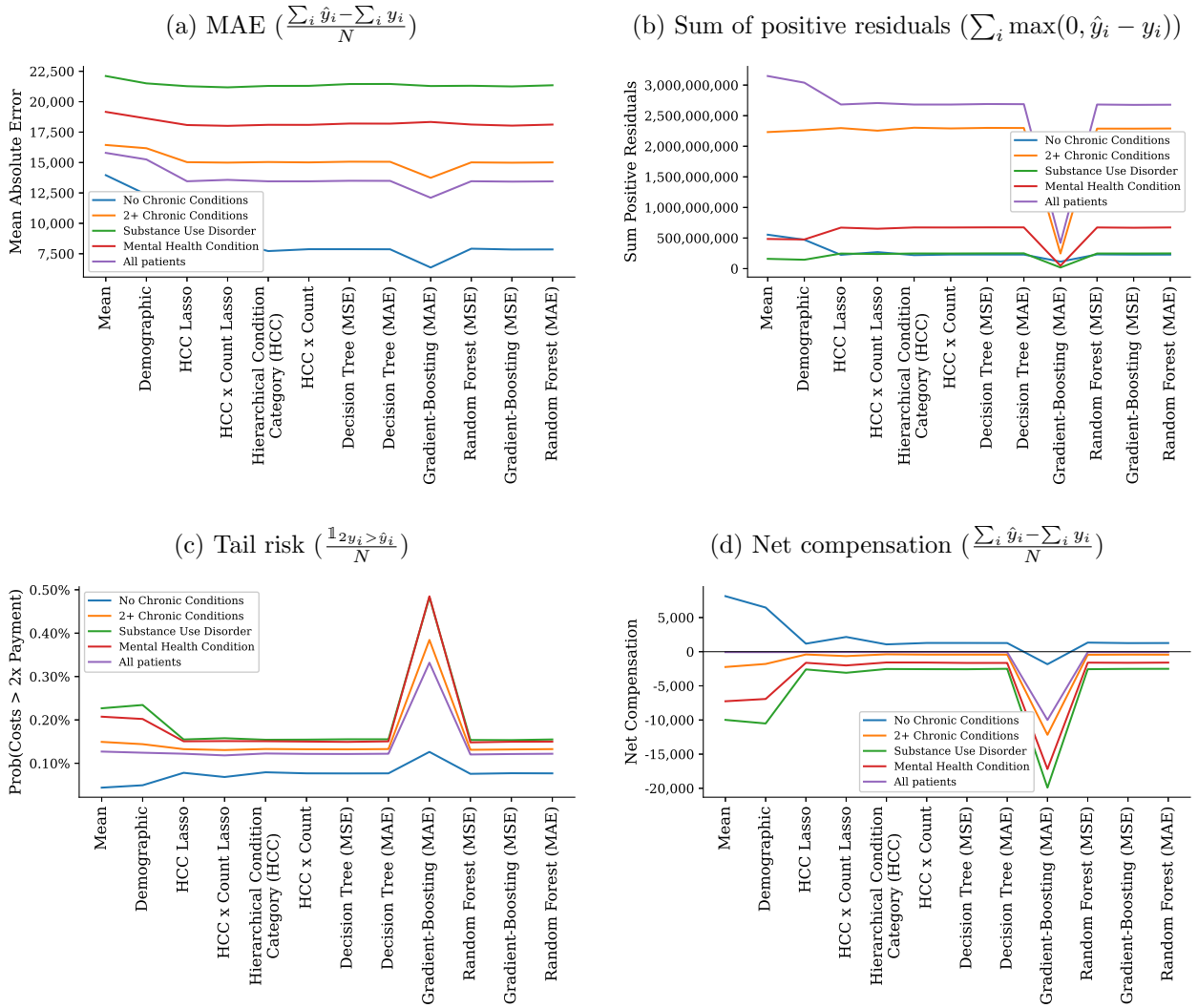
Notes: This figure restricts attention to models that are on the Pareto frontier of the mean squared error (MSE) and complexity. It is analogous to Figure but uses MSE rather than MAE. The x-axis lists model pairs in order of increasing complexity, and the y-axis shows the marginal decrease in the MSE per additional model coefficient for each pair of models. The dotted horizontal gray line shows the marginal decrease in the MSE per additional model coefficient for past risk adjustment model changes, specifically the change from the Demographic to the HCC model. I assume that model changes that offer less than 10% of this decrease are not acceptable to policymakers. As such, none of the models that are more complex than the HCC model are worth their additional complexity.

Figure D.3: Pareto Frontier of Accuracy (MSE) and Complexity (Number of Coefficients)



Notes: This graph shows a scatterplot of the accuracy and complexity of different risk adjustment models. It is analogous to figure but shows MSE rather than MAE. The x-axis, in log scale, shows model complexity, measured as the number of coefficients. The y-axis shows model accuracy as measured by the mean squared error (MSE) out of sample. The black line shows the Pareto frontier of accuracy and complexity. The improving direction is toward the origin, or left and down, toward zero MSE and zero complexity.

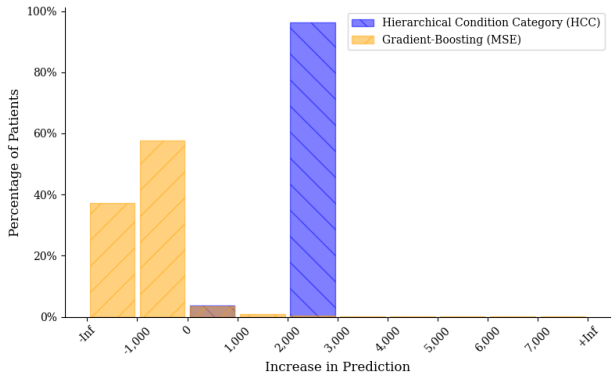
Figure D.4: Alternative Measures of Model Performance and Selection Incentives by Model and Patient Subgroup



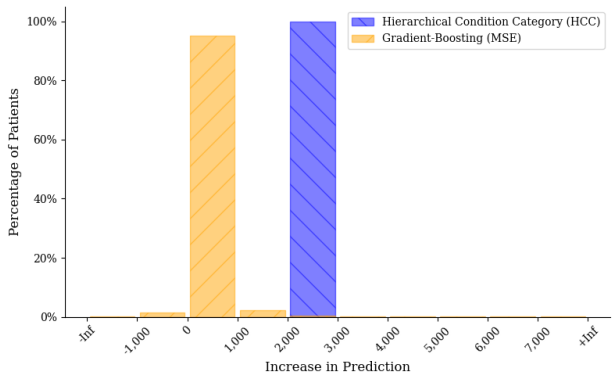
Notes: Panel (a) shows MAE on the y-axis for specified patient groups by risk adjustment model. Groups include patients with 0 chronic conditions, 2+ chronic conditions, mental health disorders, and substance use disorders, as identified by the Medicare Master Beneficiary Summary File Chronic Conditions and Other Conditions Files. These variables are based on multiple prior years of claims diagnosis data and prescription data, unlike the HCC variables, which are based only on one prior year of claims diagnosis data. Risk adjustment models are ordered on the x-axis by increasing complexity. Panel (b) is the same but with the sum of positive residuals on the y-axis. Panel (c) shows the tail risk, or the probability that realized expenditures substantially (2x) exceeds predicted spending and, therefore, payments. Panel (d) is similar but shows net compensation for the specified patient subgroups on the y-axis.

Figure D.5: Increase in Predicted Patient Cost from Upcoding

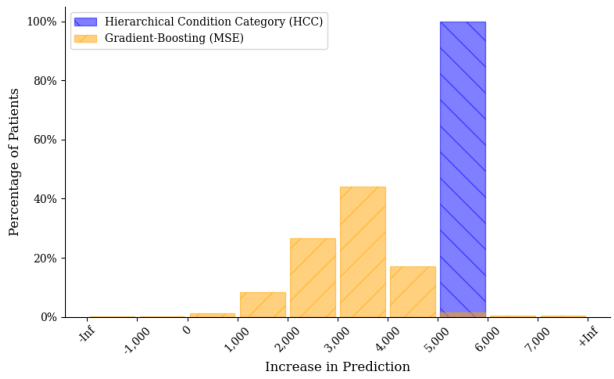
(a) Diabetes with Chronic Conditions (HCC 19)



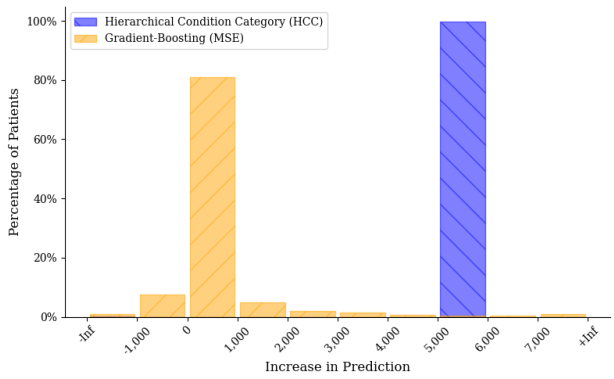
(b) Morbid Obesity (HCC 22)



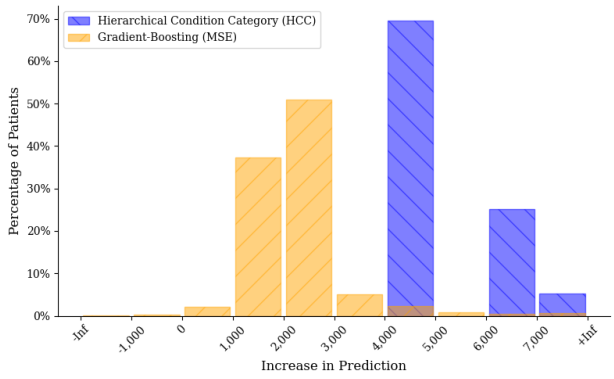
(c) Rheumatoid Arthritis (HCC 40)



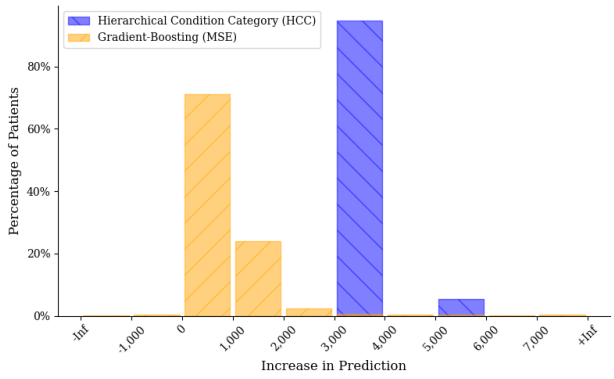
(d) Coagulation Defects (HCC 48)



(e) Congestive Heart Failure (HCC 85)



(f) Specified Heart Arrhythmias (HCC 96)



Notes:
Panel (a) shows the distribution of the change in predicted patient costs from adding diabetes with chronic complications (HCC 19) to patient records without HCC 19. It is analogous to Figure (a) but shows results for a gradient-boosted tree trained to minimize MSE, rather than MAE. 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. 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. Panels (b), (c), (d), (e), and (f) show the same analysis for adding morbid obesity (HCC 22), rheumatoid arthritis (HCC 40), coagulation defects (HCC 48), congestive heart failure (HCC 85), and heart arrhythmias (HCC 96), respectively.