Journal of Science, Technology and Social Transformation Volume 1, Issue 04, 2025

Interpretable and Fair Machine Learning for Health-Cost Prediction and Risk Adjustment: A Reproducible Benchmark on Public Data and an Open-Source Toolkit

Temitope Ologunbaba[1], Eniola Zainab Olamilekan[2], Oluwatosin Dorcas Babalola[2*], Chisom G. Adiegwu[2]

[1]Independent Researcher

[2]Maurice R. Greenberg School of Risk Science, J. Mack Robinson College of Business, Georgia State University, Atlanta, GA, USA

Abstract

Background. Risk-adjustment formulas set what health plans are paid for each enrollee. A formula that underpays a group gives plans a reason to avoid it, one that responds to recorded diagnoses rewards coding, and one that responds to past use rewards use. Machine-learning methods promise accuracy, but these three properties are rarely evaluated together, on public data, with released code.

Methods. We built a prospective benchmark from five two-year Medical Expenditure Panel Survey panels (43,963 persons, 2018 to 2023) in which year-1 information predicts year-2 total spending. Ten models, including a non-negative least-squares model in the form of a payment formula, gradient boosting and a combined actuarial neural network, were evaluated by cross-validation that keeps primary sampling units whole, with Rao-Wu bootstrap and paired intervals. Outcomes were accuracy, net compensation for nine groups, an accuracy-fairness frontier, SHAP drivers, and the response of payment to simulated coding and to added use.

Results. Gradient boosting reached an out-of-sample R^2 of 0.187 (95% interval 0.156 to 0.242) against 0.123 for weighted least squares, a paired gain of 0.064 (0.047 to 0.091) that disappeared without prior use and spending. Every model underpaid people who need help with daily activities, by $4,545 to $9,583 per person-year on a mean cost of $31,329. A constrained stacked estimator cut the largest target-group gap to $296 at an R^2 of 0.171, but both constrained estimators moved $2,242 to $3,362 of overpayment onto enrollees aged 65 and over, and out of fold the constrained gap swung between folds by more than $17,000. A coder targeting each model's own most lucrative codes raised payment by $11,045 per added code under the payment-form model and $17,469 under the constrained linear formula, against $5,510 under boosting, on the feature set without prior use.

Conclusions. Flexible models are more accurate only when they use prior utilization, which payment formulas exclude. No model pays for functional need unless it observes it. Fairness constraints hold in training but not reliably out of sample, and they move money between groups, so they must be reported for all groups with intervals.

Keywords. risk adjustment; machine learning; algorithmic fairness; net compensation; coding intensity; SHAP; MEPS

Journal of Science, Technology and Social Transformation (2025) Doi: 10.64235/184sde0e

Introduction

Risk adjustment decides how several hundred billion dollars a year are divided among health plans. In Medicare Advantage, plan payments are scaled by a risk score built from enrollee demographics and diagnoses; in the Affordable Care Act individual and small-group markets, risk scores drive transfers between insurers (Pope et al., 2004; Kautter et al., 2014). The formulas behind these scores are additive regressions of annual spending on demographic cells and hierarchical condition categories, and prospective formulas of this kind

Corresponding Author: Oluwatosin Dorcas Babalola, Maurice R. Greenberg School of Risk Science, J. Mack Robinson College of Business, Georgia State University, Atlanta, GA, USA

How to cite this article: Ologunbaba, T., Olamilekan, E.Z., Babalola, O.D., Adiegwu, C.G. (2025). Interpretable and Fair Machine Learning for Health-Cost Prediction and Risk Adjustment: A Reproducible Benchmark on Public Data and an Open-Source Toolkit. *Journal of Science, Technology and Social Transformation* 1(4), 64-82.

Source of support: Nil

Conflict of interest: None

explain a minority of the variation in individual spending (Ellis, Martins and Rose, 2018).

Three problems with these formulas are well documented and usually studied one at a time. First, they leave predictable variation unused, and flexible methods can recover some of it (Rose, 2016; Duncan, Loginov and Ludkovski, 2016; Irvin et al., 2020). Second, they underpay identifiable groups, and a plan that can identify an underpaid group profits from avoiding it: Brown et al. (2014) found that Medicare Advantage plans selected on the risk left unpriced after risk adjustment was refined, and Montz et al. (2016) showed that marketplace plans had incentives to distort mental health and substance use coverage. Constrained and penalized regressions can remove chosen underpayments at a cost in fit (van Kleef et al., 2017; Zink and Rose, 2020; McGuire, Zink and Rose, 2021). Third, payment follows what is recorded. Risk scores rise when the same people move into Medicare Advantage (Geruso and Layton, 2020), coding intensity in Medicare Advantage has been measured for more than a decade (Kronick and Welch, 2014), and risk scores there have risen faster than other measures of beneficiary health would imply (Jacobs and Kronick, 2018). The Medicare Payment Advisory Commission projected that Medicare would spend 20% more on Medicare Advantage enrollees in 2025 than if they were in fee-for-service Medicare, about $84 billion; it attributed $40 billion to coding intensity, with risk scores about 16% above those of comparable fee-for-service beneficiaries before a 5.9% statutory adjustment, and $44 billion to favorable selection (MedPAC, 2025). CMS's V28 model, finalized in the 2024 Rate Announcement and phased in over 2024 to 2026, rebuilt the condition categories from ICD-10 codes, cut the diagnosis codes that map to payment categories from 9,797 to 7,770, and dropped some conditions whose diagnosis and coding CMS judged discretionary (CMS, 2023).

A fourth channel is less discussed. A formula that uses prior spending or prior use pays more next year for more care this year. Payment formulas exclude prior use for that reason, because a payment that reimburses a provider's or plan's own use gives up the incentive to control it (Ellis and McGuire, 1986, 1993), and use responds to financial incentives in ways that differ by type of service (Ellis, Martins and Zhu, 2017). Many machine-learning cost models include prior use because it is the strongest predictor.

This paper evaluates all four properties together, on public data, with released code. Using five two-year panels of the Medical Expenditure Panel Survey (MEPS), we predict each person's next-year spending from current-year information and compare ten models: unconstrained weighted least squares, a non-negative least-squares model in the form payment formulas take, a Tweedie generalized linear model, a two-part model, an elastic net, gradient boosting with two objectives, a random forest, a feed-forward network and a combined actuarial neural network. We ask five questions.

- RQ1. How much more accurately do flexible models predict next-year spending than the linear payment form, where do the gains come from, and do they survive without prior use?
- RQ2. Which groups does each model under- or overpay, with what uncertainty, and how does each model differ from the linear form on the same resamples?
- RQ3. How much accuracy does it cost to close the gaps for chosen groups, what happens to the other groups, and does the closure hold out of sample?
- RQ4. What drives the flexible models' predictions, and are the drivers stable and sensible?
- RQ5. How much does each model's payment rise when coding intensifies, at random, toward chronic conditions, toward the codes a published formula pays most, and toward the codes that model pays most, and how much does it rise per added dollar of prior use?

Three design choices distinguish the benchmark. Cross-validation keeps MEPS primary sampling units whole and intervals come from a bootstrap that respects the survey design, so the accuracy of a model that memorizes local effects is not overstated. Race and ethnicity never enter a model; they are used only to evaluate what a model does. And every comparison between two models is paired on the same resamples, so a reader can tell a difference from noise.

Relative to earlier machine-learning risk adjustment (Rose, 2016; Shrestha et al., 2018; Irvin et al., 2020; Andriola et al., 2024), the contribution is not a new estimator but a joint evaluation: the same models, the same folds and the same groups for accuracy, fairness, stability of fairness out of sample, coding response and use response, released as a package that runs on any person-level cost data.

Background

Payment formulas

The CMS-HCC and HHS-HCC models regress annualized spending on age-sex cells and hierarchical condition categories derived from diagnosis codes. They have no free intercept, their demographic cells are sex-specific, and condition coefficients that would be negative or would reverse a clinical hierarchy are constrained (Pope et al., 2004; Kautter et al., 2014). People enrolled for part of a year enter with spending annualized and a weight equal to their fraction of the year. The additive form is chosen for transparency, ease of administration and resistance to manipulation. Ellis, Martins and Rose (2018) review the design choices, Layton et al. (2017) set out how fit and incentives should be weighed against each other, and Park and Basu (2018) propose evaluation metrics beyond R^2.

Machine learning for cost prediction

Rose (2016) compared a library of machine-learning estimators for plan payment on commercial claims; an ensemble performed best, and a simplified formula with fewer variables kept much of the efficiency of the full one. Duncan, Loginov and Ludkovski (2016) compared regression

frameworks including boosting for cost prediction. Shrestha et al. (2018) applied machine learning to mental health risk adjustment. Irvin et al. (2020) added social determinants. Andriola et al. (2024) introduced a learning algorithm that builds payment formulas within the constraints of a payment system. Wüthrich and Merz (2023) give the actuarial foundations of the models used here, and Schelldorfer and Wüthrich (2019) the combined actuarial neural network, in which a GLM's linear predictor is a fixed skip connection and a network learns a correction.

Fairness

Plan-payment research measures fairness at the group level: whether a formula pays for a group what the group costs. Van Kleef et al. (2017) used constrained regression to remove the undercompensation of chosen groups in the Dutch system. Zink and Rose (2020) proposed fair regression estimators that penalize or constrain net compensation, the difference between what a formula pays for a group and what the group costs, and McGuire, Zink and Rose (2021) compared constrained regression with reinsurance and variable selection. Zink and Rose (2021) showed that undercompensated groups can be defined by combinations of attributes that single-attribute checks miss. Bergquist et al. (2019) showed that transforming the data used to estimate a formula can improve its performance for undercompensated groups without changing the formula's form. Obermeyer et al. (2019) showed that an algorithm trained on cost as a proxy for need underrated Black patients, which is why the outcome here is spending and the interpretation is payment, not need.

Coding and use

Coding intensity is a response to the formula: codes that pay are recorded more often. Geruso and Layton (2020) and MedPAC (2025) measure the result in Medicare Advantage. The sensitivity of a formula to added codes, and to added use when prior use is a predictor, is a property of the formula that can be measured before any behavior changes, and Section 6.6 does so.

Data

Panels

MEPS follows each panel of households for two calendar years. We used the two-year longitudinal files for panels 23 to 27, covering 2018-2019 through 2022-2023, and the Medical Conditions file for each panel's first year. Panels 23 and 24 were later extended; we used their original two-year files so that every person contributes one prospective pair with the two-year longitudinal weight.

Sample

A person enters if present in year 1 with a positive longitudinal weight (Table A1). The 395 people who died in year 2 are kept, as payment models keep partial-year enrollees: their year-2 spending is divided by the fraction of year 2 they were in scope, and their weight is multiplied by the same fraction, so each contributes in proportion to exposure and weighted total spending is unchanged. Robustness checks drop them and use unannualized spending. The analysis sample is 43,963 persons representing 325.8 million a year.

Outcome

Year-2 total expenditure from all payers, in 2024 dollars using the CPI-U. The weighted mean is $7,792, the median $1,497, 14.9% spent nothing, the 99th percentile is $95,759, and the top 5% of spenders account for 50.7% of spending (Table 1).

Predictors

All from year 1. Demographics: age band by sex and Census region. Conditions: AHRQ Clinical Classifications Software Refined (CCSR) categories from the conditions file, as binary flags for the 124 categories held by at least 0.5% of the sample (Table A6), plus counts of categories and body systems. Two CCSR categories, diabetes without complication (END002) and type 2 diabetes (END005), are assigned from the same three-digit codes and are identical in MEPS; they enter as one flag. MEPS assigns body-system placeholder codes (XXX000) where a diagnosis cannot be placed in a category; these are kept in the body-system count and excluded from flags and condition counts. Use: office visits, emergency visits, inpatient discharges, prescription fills and total spending, logged. Perceived health and mental health are taken from round 2 and help with activities of daily living (ADL) and instrumental activities (IADL) from round 1, because round 3 is interviewed in year 2 and would leak year-2 information.

Four nested feature sets are used

F1 demographics; F2 adds conditions; F3 adds prior use and spending; F4 adds income, insurance, perceived health and functional help. F3 is the primary set because it is what cost-prediction studies usually use. F2 is the set closest to a deployable payment formula and is reported for every result where the difference matters.

Groups for evaluation

Income below 200% of the federal poverty level; Hispanic, non-Hispanic Black and non-Hispanic Asian; needs help with ADL or IADL; any mental or behavioral condition (CCSR MBD); age 65 and over; uninsured all year; and died in year 2. Race and ethnicity are never model inputs. Three groups are targets for the fairness estimators: low income, mental health conditions and functional help. They were chosen because the payment literature identifies them as undercompensated and because a payer can observe them. The ADL group has 1,612 persons, 2.8% of the weighted population.

CCSR limitation

MEPS releases three-digit ICD-10-CM codes. CCSR categories built from them are coarser than the hierarchical condition categories of the CMS and HHS models, so the payment-form

Table 1: Sample characteristics and year-2 spending

Characteristic	*Estimate*	*SE*	*n*
Persons	43,963.00		43,963
Weighted population, millions a year	325.79		43,963
Age in year 1, mean	38.92	0.20	43,963
Age 65 and over, %	16.47	0.27	43,963
Female, %	50.91	0.24	43,963
Hispanic, %	19.01	0.60	43,963
Non-Hispanic White, %	58.64	0.64	43,963
Non-Hispanic Black, %	12.44	0.42	43,963
Non-Hispanic Asian, %	6.15	0.30	43,963
Non-Hispanic other or multiple, %	3.76	0.19	43,963
Below 100% FPL, %	12.36	0.34	43,963
100-124% FPL, %	4.07	0.20	43,963
125-199% FPL, %	12.28	0.34	43,963
200-399% FPL, %	28.91	0.48	43,963
400% FPL or more, %	42.38	0.58	43,963
Any private coverage, %	65.82	0.50	43,963
Public coverage only, %	27.49	0.44	43,963
Uninsured all year, %	6.68	0.23	43,963
Fair or poor perceived health, %	9.78	0.21	43,789
Needs help with ADL or IADL, %	2.82	0.10	43,963
Any year-1 condition (CCSR category), %	70.04	0.39	43,963
Year-1 CCSR categories, mean	2.80	0.03	43,963
Any mental health condition, %	15.81	0.28	43,963
Year-1 spending, mean $	7,389.49	158.11	43,963
Year-2 spending, mean $	7,792.21	146.60	43,963
Year-2 spending of zero, %	14.94	0.32	43,963
Year-2 spending, median $	1,497.40		43,963
Year-2 spending, 90th percentile $	18,542.65		43,963
Year-2 spending, 99th percentile $	95,758.64		43,963
Share of year-2 spending by the top 5% of spenders, %	50.75		43,963

Weighted estimates; standard errors Taylor-linearized for the stratified cluster design.

model here has the form of a payment formula without being any specific formula.

Methods

Models

All models share one interface and are fitted with the MEPS longitudinal weight.

Linear forms. WLS is unconstrained weighted least squares on the features with an intercept. The payment-form model is weighted least squares in the form of the CMS and HHS formulas: sex-specific age cells in place of an intercept, condition flags, counts and (on F3) use terms, every coefficient constrained to be non-negative, no region term, solved by bounded least squares. It is not rescaled, so its predictions need not balance to total spending; a payment system would apply a normalization factor. Unconstrained WLS is reported because it is what "linear" usually means in cost prediction; it predicts a negative payment for 15.2% of the weighted population, which no payment system could use.

Other models. The Tweedie GLM has a log link and variance power 1.5. The two-part model multiplies a logistic probability of any spending by a gamma GLM for the amount. The elastic net is fitted on raw spending. Gradient boosting uses LightGBM with a Tweedie objective and, separately, a squared-error objective. The random forest uses 300 trees. The neural network has two hidden layers, a log link and Tweedie deviance loss. The combined actuarial

neural network (CANN) adds the fitted Tweedie GLM's linear predictor as a fixed offset and initializes the network's output layer at zero, so it starts at the GLM and moves only where validation deviance improves. Multiplicative models are rescaled on their own training data so that weighted predictions sum to weighted spending.

Every hyperparameter is chosen inside the training data of each outer fold. The GLM, two-part and elastic-net penalties are chosen by three inner cluster-disjoint folds; boosting and the networks use one inner split of whole clusters for early stopping and a small grid. No test observation influences any choice (Table A5).

Survey-aware cross-validation and intervals

Strata are MEPS variance strata within panel. Within each stratum, primary sampling units (PSUs) are shuffled and dealt in turn across five folds, so no PSU contributes to both a training and a test set. Because most strata hold only a few PSUs, a fold does not contain every stratum; the deal balances folds across strata as far as the design allows. The procedure is repeated three times with different assignments.

Each fitted model predicts its held-out PSUs, so every person receives one out-of-fold prediction per repeat. The headline metric is the mean over the three repeats of each repeat's metric, which is what one deployed model achieves. Averaging the three repeats' predictions first is a small ensemble; its R^2 is reported separately (Table 2) and is larger mainly for the neural network.

Intervals come from the Rao-Wu rescaled bootstrap (Rao and Wu, 1988): within each stratum with n PSUs, n − 1 PSUs are drawn with replacement and the weights of drawn persons are multiplied by n/(n − 1) times the number of times their PSU was drawn. This keeps the variance of a design with few PSUs per stratum from being understated, which a naive PSU bootstrap does. Accuracy, group and frontier intervals use 200 resamples. Paired differences between models are computed on resamples shared by both models, 500 for accuracy and 200 for group net compensation, so the interval is for the difference itself. The bootstrap holds the fitted predictions fixed and so reflects the sampling of the evaluation population, not refitting; the spread of R^2 across repeats, which does reflect refitting, is at most 0.005 for every model except the neural network (0.015).

Metrics

Accuracy: weighted R^2, mean absolute error, Cumming's prediction measure (CPM; Cumming et al., 2002), the absolute-error analogue of R^2, predictive ratios by decile of predicted spending and by the age-sex cells of a payment formula, the share of the true top 10% of spenders placed in the predicted top 10%, and the share of negative predictions. Fairness, following Zink and Rose (2020): for group G, the predictive ratio PR_G is weighted predicted over weighted observed spending, and net compensation NC_G is the weighted mean of prediction minus outcome, in dollars per person-year. Negative NC_G means the group is underpaid.

Fairness estimators

Because NC_G is linear in the coefficients of a linear model, two estimators have exact solutions. Penalized WLS adds $\lambda \Sigma_G NC_G^2$ to the weighted squared-error loss, with λ scaled by the outcome variance so it is unit-free; constrained WLS imposes NC_G = 0 for each target group, the $\lambda \to \infty$ limit, solved as a KKT system. The grid is λ = 0, 0.1, 0.3, 1, 3, 10, 30 and 100 and the constrained limit.

For boosting we use a stacked form. A LightGBM score is cross-fitted within the training data (three cluster-disjoint folds, so no training row's score was fitted to its own outcome), and the score enters a penalized or constrained regression together with the features. We tried adding the net-compensation penalty inside the boosting objective and rejected it: the penalty's curvature is shared across a group, a diagonal Hessian either understates it, and Newton steps diverge, or, if spread across members, damps the fitting of their own residuals, and in testing underpayment of small high-cost groups worsened as the penalty rose. The stacked form keeps the boosted model's accuracy in the score and puts the adjustment where it has an exact solution.

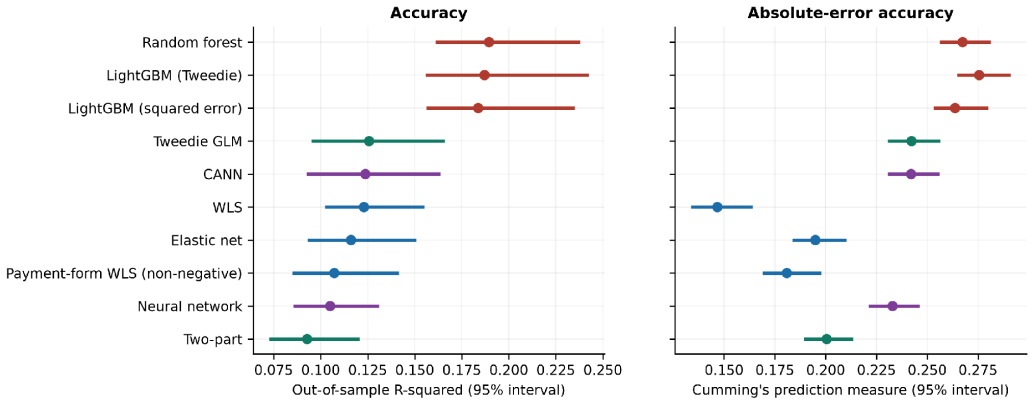


Figure 1: Out-of-sample R^2 and Cumming's prediction measure by model, primary feature set, with 95% Rao-Wu bootstrap intervals.

Table 2: Out-of-sample accuracy by model, primary feature set (F3)

Model	R^2	*95% low*	*95% high*	*R^2, 3-fit ensemble*	*CPM*	*MAE*	*PR bottom decile*	*PR top decile*	*Top-10% capture*	*Negative predictions*
WLS	0.123	0.102	0.155	0.124	0.147	$8,337	-1.94	0.94	44%	15.2%
Payment-form WLS (non-negative)	0.107	0.085	0.142	0.108	0.181	$8,003	0.08	1.01	41%	0.0%
Tweedie GLM	0.126	0.095	0.166	0.130	0.242	$7,403	0.98	1.00	47%	0.0%
Two-part	0.093	0.073	0.121	0.099	0.201	$7,811	1.00	0.96	43%	0.0%
Elastic net	0.116	0.093	0.151	0.117	0.195	$7,865	-0.83	0.94	42%	12.3%
LightGBM (Tweedie)	0.187	0.156	0.242	0.190	0.276	$7,077	1.03	0.99	48%	0.0%
LightGBM (squared error)	0.184	0.156	0.235	0.189	0.264	$7,193	1.43	1.00	47%	0.0%
Random forest	0.190	0.161	0.238	0.191	0.267	$7,157	0.75	0.98	48%	0.0%
Neural network	0.105	0.085	0.131	0.135	0.233	$7,494	0.98	1.05	42%	0.0%
CANN	0.124	0.093	0.164	0.129	0.242	$7,405	0.95	1.01	47%	0.0%

Five survey folds, three repeats. Each metric is the mean over repeats of a single fitted model's out-of-fold value, which is what one deployed model achieves; the ensemble column averages the three repeats' predictions first. Intervals from a Rao-Wu rescaled bootstrap over PSUs within strata (200 resamples). CPM is Cumming's prediction measure. Top-10% capture is the weighted share of the true top decile of spenders placed in the model's top decile. The last column is the weighted share of people with a negative predicted payment.

As a group-neutral comparison, the Tweedie LightGBM's predictions are recalibrated by decile of predicted spending, with factors estimated on other folds' out-of-fold predictions.

The frontier is run on all three repeats of the folds, with Rao-Wu intervals for R^2 and for the largest absolute target-group gap. Because a constraint that holds exactly in training need not hold on new data, we also report, for each of the 15 folds, the target groups' net compensation in the training data and out of fold (Table 5b). Group membership enters estimation only through group means and never enters the prediction formula.

Drivers

TreeSHAP values are computed for the squared-error LightGBM on up to 4,000 held-out persons per fold, from the model fitted on that fold's training PSUs. The squared-error model is explained because its SHAP values are in dollars, and its accuracy is within 0.003 of the Tweedie model's. Global importance is the weighted mean absolute SHAP value; stability is the Spearman correlation of importance ranks between folds. For the CANN, permutation importance on the held-out fold is used. Sense checks test whether SHAP contributions rise with year-1 spending, the number of conditions and the number of body systems.

Coding and use sensitivity

Models are fitted on the training PSUs of each fold of the first repeat and stressed on that fold's held-out PSUs, on F3 and on F2. For a share of held-out persons (5%, 10% or 25%), one or two condition flags they did not have are switched on, and the condition and body-system counts are updated; use and spending are unchanged. The flags come from one of four pools:

- prevalence: drawn in proportion to each category's prevalence among people with any condition;
- chronic: drawn from payment-relevant chronic categories (diabetes, coronary disease, dysrhythmia, heart failure, COPD, depression and others listed in the code; six are common enough to be features);
- targeted: the five flags with the largest coefficients in the payment-form model, which is what a coder who knows a published formula adds (polyneuropathies, rheumatoid arthritis, other kidney disease, epilepsy and heart failure, on both feature sets);
- own-targeted: for each model and fold separately, the five flags that raise that model's own prediction most when switched on for 3,000 training persons, which is what a coder who can query the model adds (Table 7c).

We report the rise in predicted spending per added code, pooled across folds, with the standard deviation across folds.

For use sensitivity, year-1 spending is raised by 20% for 10% of held-out persons with nothing else changed, and we record the rise in next-year predicted spending per added dollar of year-1 spending. Visits, fills and discharges are held fixed, so the experiment isolates the spending term; a real rise in use would move those counts too.

Robustness

Each variant changes one decision and reruns WLS, the payment-form model, the Tweedie GLM and Tweedie LightGBM on one repeat: outcome capped at $250,000; a log-outcome linear model; unweighted training; condition-flag prevalence floors of 0.2% and 1%; feature sets F2 and F4; decedents excluded; decedents' spending not annualized; three and ten folds; each panel alone; panels whose outcome

Table 2b: Paired differences from WLS in out-of-sample accuracy

Features	*Model*	*ΔR²*	*95% low*	*95% high*	*Share ΔR² ≤ 0*	*ΔCPM*	*ΔMAE*	*95% low*	*95% high*
F3	Payment-form WLS (non-negative)	-0.016	-0.020	-0.010	1.000	0.034	-$334	-$372	-$295
F3	Tweedie GLM	0.003	-0.018	0.025	0.532	0.096	-$934	-$995	-$879
F3	Two-part	-0.030	-0.042	-0.020	1.000	0.054	-$526	-$578	-$469
F3	Elastic net	-0.007	-0.011	-0.001	0.988	0.048	-$472	-$502	-$442
F3	LightGBM (Tweedie)	0.064	0.047	0.091	0.000	0.129	-$1,259	-$1,327	-$1,196
F3	LightGBM (squared error)	0.061	0.044	0.083	0.000	0.117	-$1,143	-$1,213	-$1,077
F3	Random forest	0.067	0.053	0.085	0.000	0.121	-$1,180	-$1,243	-$1,119
F3	Neural network	-0.018	-0.038	-0.004	0.998	0.086	-$843	-$894	-$792
F3	CANN	0.001	-0.020	0.023	0.544	0.095	-$932	-$992	-$877
F1	LightGBM (Tweedie)	-0.001	-0.003	0.000	0.950	-0.004	$37	$22	$53
F1	Payment-form WLS (non-negative)	-0.000	-0.001	0.000	0.920	-0.000	$4	-$8	$17
F2	LightGBM (Tweedie)	-0.002	-0.006	0.003	0.772	0.012	-$116	-$151	-$85
F2	Payment-form WLS (non-negative)	-0.005	-0.008	-0.001	0.994	0.004	-$39	-$70	-$12
F4	LightGBM (Tweedie)	0.061	0.046	0.085	0.000	0.128	-$1,247	-$1,312	-$1,181
F4	Payment-form WLS (non-negative)	-0.014	-0.018	-0.008	1.000	0.033	-$323	-$360	-$283

Every model is evaluated on the same 500 resamples of PSUs within strata, so the interval is for the difference itself. The last column is the share of resamples in which the model's R^2 does not exceed WLS's. Negative MAE differences mean smaller errors than WLS.

year fell in 2020 or 2021 dropped; and the subsets aged 65 and over and under 65 with private coverage.

Software

The estimators, folds, bootstrap, metrics, fairness methods and coding stress test are released as `riskfair`, a Python package with no dependence on this paper's configuration: every choice is an argument, so it runs on any person-level cost data with weights and design identifiers, survey or claims. The pipeline that produces every table and figure calls the package. Unit tests cover fold disjointness, the Rao-Wu weights, metric identities, exact zero net compensation for the constrained estimator in training, monotone response to the penalty, recovery of the constraint by the stacked estimator on simulated data, and the coding perturbation and per-flag gains. A test asserts that no feature name refers to race or to the year-2 outcome and that no feature is nearly collinear with the outcome. A script checks every number quoted in this text against the output tables.

Results

Accuracy (RQ1)

Table 2 and Figure 1 give out-of-sample accuracy on the primary feature set. The three tree ensembles reach R^2 of 0.184 to 0.190: Tweedie LightGBM 0.187 (0.156 to 0.242), squared-error LightGBM 0.184 and the random forest 0.190. WLS reaches 0.123 (0.102 to 0.155) and the payment-form model 0.107 (0.085 to 0.142). The Tweedie GLM reaches 0.126, the CANN 0.124, the elastic net 0.116, the neural network 0.105 and the two-part model 0.093. Given the same features, the network correction in the CANN found little to add to its GLM base. The neural network's R^2 rises to 0.135 when its three repeats are averaged, the only model for which the ensemble matters.

Table 2b pairs every model with WLS on the same resamples. Tweedie LightGBM gains 0.064 (0.047 to 0.091), the random forest 0.067 (0.053 to 0.085) and squared-error LightGBM 0.061; no resample favors WLS. The Tweedie GLM (0.003, −0.018 to 0.025) and the CANN are indistinguishable from WLS on R^2, while the neural network (−0.018, −0.038 to −0.004) and the payment-form model (−0.016, −0.020 to −0.010) are worse. On absolute error every model beats WLS, because WLS fits the mean of a skewed outcome with negative predictions at the bottom: its predictive ratio in the bottom decile of predicted spending is −1.94. Cumming's prediction measure rises from 0.147 for WLS to 0.181 for the payment-form model, 0.242 for the Tweedie GLM and 0.276 for Tweedie LightGBM.

Table 3: Out-of-sample R^2 by feature set

Model	*F1*	*F2*	*F3*	*F4*
LightGBM (Tweedie)	0.039	0.101	0.187	0.189
Payment-form WLS (non-negative)	0.040	0.098	0.107	0.113
Tweedie GLM	0.041	-0.001	0.126	0.123
WLS	0.040	0.103	0.123	0.127

F1 demographics; F2 adds year-1 CCSR condition flags; F3 adds year-1 use and spending; F4 adds income, insurance, perceived health and functional help.

Table 4: Net compensation by group: predicted minus observed spending, $ per person-year (95% interval), and predictive ratio

Group	*WLS*	*Payment-form WLS (non-negative)*	*Tweedie GLM*	*LightGBM (Tweedie)*	*CANN*
Age 65 and over	-$1 (-$903, $686); PR 1.00	$189 (-$681, $882); PR 1.01	$562 (-$309, $1,297); PR 1.03	-$467 (-$1,298, $176); PR 0.97	$645 (-$232, $1,371); PR 1.04
Died in year 2	-$43,607 (-$54,318, -$32,699); PR 0.34	-$44,316 (-$55,359, -$33,097); PR 0.33	-$42,318 (-$52,745, -$31,763); PR 0.36	-$42,541 (-$53,195, -$31,878); PR 0.36	-$42,174 (-$52,655, -$31,638); PR 0.36
Hispanic	-$271 (-$1,034, $336); PR 0.94	$509 (-$297, $1,114); PR 1.11	$12 (-$777, $609); PR 1.00	$91 (-$670, $697); PR 1.02	-$0 (-$794, $594); PR 1.00
Income below 200% FPL	-$92 (-$619, $336); PR 0.99	$675 ($124, $1,113); PR 1.09	$133 (-$391, $565); PR 1.02	-$5 (-$514, $417); PR 1.00	$149 (-$377, $580); PR 1.02
Mental health condition	$193 (-$474, $772); PR 1.01	$312 (-$363, $912); PR 1.02	$675 (-$11, $1,367); PR 1.05	$538 (-$38, $1,128); PR 1.04	$724 ($41, $1,418); PR 1.05
Needs ADL or IADL help	-$7,747 (-$12,747, -$4,246); PR 0.75	-$9,293 (-$14,437, -$5,593); PR 0.70	-$4,762 (-$9,353, -$1,287); PR 0.85	-$5,721 (-$10,502, -$2,458); PR 0.82	-$4,545 (-$9,069, -$1,057); PR 0.85
Non-Hispanic Asian	$179 (-$399, $735); PR 1.04	$780 ($225, $1,315); PR 1.16	$387 (-$152, $867); PR 1.08	$310 (-$237, $838); PR 1.06	$370 (-$172, $850); PR 1.08
Non-Hispanic Black	-$202 (-$896, $485); PR 0.97	$413 (-$230, $1,096); PR 1.06	-$15 (-$680, $629); PR 1.00	-$8 (-$645, $603); PR 1.00	-$16 (-$681, $620); PR 1.00
Uninsured all year	-$126 (-$696, $373); PR 0.95	$657 ($102, $1,197); PR 1.27	$425 (-$118, $925); PR 1.18	$549 ($40, $1,024); PR 1.23	$408 (-$134, $906); PR 1.17

Negative values mean the model pays less for the group than the group costs. Means over repeats of single fits; Rao-Wu PSU bootstrap intervals. Race and ethnicity are used for evaluation only and never as model inputs.

The boosting gain sits where payment formulas are weakest. The predictive ratio in the top decile of predicted spending is 0.94 for WLS and 0.99 for Tweedie LightGBM (Figure 2 and Table A2), the top-10% capture rate rises from 44% to 48%, and mean absolute error falls from $8,337 to $7,077.

Table 3 shows where the information is. On demographics alone every model reaches R^2 of about 0.040. Adding conditions raises WLS to 0.103, the payment-form model to 0.098 and boosting to 0.101; adding prior use and spending raises WLS to 0.123 and boosting to 0.187. The boosting advantage therefore comes from how it uses prior use, which payment formulas exclude by design. On F2, the set closest to a deployable formula, the paired difference between

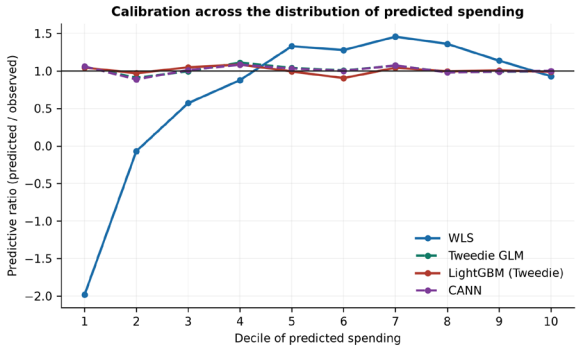


Figure 2: Predictive ratio by decile of predicted spending, four models

Table 4b: Difference in net compensation from WLS, by group (95% paired interval)

Group	*Payment-form WLS (non-negative)*	*Tweedie GLM*	*LightGBM (Tweedie)*	*CANN*
Age 65 and over	$191 ($75, $283)	$563 ($314, $871)	-$465 (-$680, -$272)	$647 ($395, $956)
Died in year 2	-$709 (-$1,403, -$157)	$1,289 ($224, $2,854)	$1,066 (-$802, $3,244)	$1,433 ($323, $3,068)
Hispanic	$780 ($689, $863)	$284 ($204, $370)	$362 ($239, $497)	$271 ($189, $360)
Income below 200% FPL	$767 ($681, $868)	$224 ($103, $374)	$87 (-$34, $253)	$241 ($116, $398)
Mental health condition	$118 ($17, $241)	$481 ($194, $795)	$344 ($119, $600)	$530 ($240, $857)
Needs ADL or IADL help	-$1,546 (-$1,949, -$1,188)	$2,985 ($2,227, $3,834)	$2,025 ($1,354, $2,925)	$3,201 ($2,439, $4,064)
Non-Hispanic Asian	$602 ($425, $757)	$209 ($40, $392)	$131 (-$54, $331)	$191 ($19, $373)
Non-Hispanic Black	$616 ($489, $739)	$187 ($42, $359)	$194 (-$9, $391)	$187 ($38, $365)
Uninsured all year	$783 ($609, $930)	$552 ($406, $688)	$675 ($491, $862)	$534 ($388, $670)

Same PSU resamples for both models, so the interval is for the difference itself. Positive values mean the model pays the group more than WLS does.

Table 5: The accuracy-fairness frontier

Method	*λ*	*R^2*	*low*	*high*	*Max*	*NC*	*target*	*low*	*high*	*Income below 200% FPL*	*Hispanic*	*Non-Hispanic Black*	*Non-Hispanic Asian*	*Needs ADL or IADL help*	*Mental health condition*	*Age 65 and over*
Penalized WLS	0	0.123	0.102	0.155	$7,747	$4,246	$12,747	-$92	-$271	-$202	$179	-$7,747	$193	-$1	-$126	-$43,607
Penalized WLS	0.1	0.119	0.099	0.150	$4,504	$1,240	$9,351	$548	-$89	$213	$383	-$4,504	$1,053	$1,690	-$95	-$40,148
Penalized WLS	0.3	0.114	0.094	0.142	$2,817	$726	$7,644	$664	-$221	$232	$316	-$2,817	$858	$2,554	-$302	-$38,145
Penalized WLS	1	0.107	0.088	0.132	$1,574	$500	$6,366	$487	-$605	-$19	$31	-$1,574	$421	$3,115	-$758	-$36,610
Penalized WLS	3	0.103	0.085	0.127	$1,044	$357	$5,843	$252	-$950	-$297	-$243	-$1,044	$185	$3,293	-$1,148	-$35,970
Penalized WLS	10	0.100	0.083	0.124	$825	$267	$5,619	$101	-$1,153	-$469	-$407	-$825	$85	$3,345	-$1,375	-$35,715
Penalized WLS	30	0.100	0.082	0.124	$758	$235	$5,550	$47	-$1,224	-$531	-$466	-$758	$55	$3,357	-$1,455	-$35,639
Penalized WLS	100	0.099	0.082	0.123	$734	$224	$5,525	$26	-$1,251	-$554	-$488	-$734	$44	$3,361	-$1,485	-$35,612
Constrained WLS	constrained	0.099	0.082	0.123	$724	$220	$5,515	$17	-$1,263	-$565	-$497	-$724	$40	$3,362	-$1,498	-$35,600
Stacked LightGBM	0	0.182	0.152	0.234	$4,601	$1,367	$9,257	$14	-$11	-$10	$328	-$4,601	$192	$94	$313	-$41,025
Stacked LightGBM	0.1	0.180	0.152	0.231	$2,415	$620	$6,942	$424	$111	$260	$464	-$2,415	$747	$1,196	$340	-$38,712
Stacked LightGBM	0.3	0.178	0.150	0.227	$1,277	$623	$5,881	$487	$21	$265	$416	-$1,277	$615	$1,749	$204	-$37,382
Stacked LightGBM	1	0.175	0.147	0.223	$536	$480	$5,111	$355	-$239	$87	$220	-$437	$327	$2,100	-$101	-$36,367
Stacked LightGBM	3	0.173	0.145	0.221	$317	$439	$4,786	$189	-$473	-$105	$33	-$78	$173	$2,205	-$364	-$35,945
Stacked LightGBM	10	0.172	0.144	0.220	$295	$391	$4,652	$84	-$610	-$224	-$79	$71	$108	$2,233	-$517	-$35,778
Stacked LightGBM	30	0.172	0.144	0.219	$288	$386	$4,611	$46	-$659	-$267	-$119	$117	$89	$2,239	-$571	-$35,728
Stacked LightGBM	100	0.171	0.143	0.219	$293	$372	$4,596	$32	-$677	-$283	-$134	$133	$82	$2,241	-$592	-$35,711
Stacked LightGBM, constrained	constrained	0.171	0.143	0.219	$296	$365	$4,590	$26	-$685	-$290	-$141	$140	$79	$2,242	-$601	-$35,703
LightGBM, decile recalibration	0	0.186	0.155	0.241	$5,608	$2,354	$10,391	$21	$111	$18	$346	-$5,608	$597	-$430	$556	-$42,449

Penalized and constrained estimators target the three groups marked in the text; the other groups show spillover. Stacked LightGBM enters a cross-fitted LightGBM score with the features into the fair regression. All three repeats of the survey folds; metrics are means over repeats of out-of-fold values, with Rao-Wu PSU bootstrap intervals for R^2 and the largest target-group gap.

Table 5b. Constrained estimators fold by fold: net compensation in the training data and out of fold

Method	*Repeat*	*Fold*	*in: Income below 200% FPL*	*out: Income below 200% FPL*	*in: Mental health condition*	*out: Mental health condition*	*in: Needs ADL or IADL help*	*out: Needs ADL or IADL help*	*in: Age 65 and over*	*out: Age 65 and over*
Constrained WLS	0	0	$0	$481	-$0	-$1,014	-$0	$2,755	$3,991	$3,040
Stacked LightGBM, constrained	0	0	$0	$528	-$0	-$723	-$0	$2,272	$2,897	$1,604
Constrained WLS	0	1	$0	$439	-$0	$1,174	-$0	-$331	$3,486	$2,240
Stacked LightGBM, constrained	0	1	$0	$557	$0	$1,598	$0	$4,589	$2,202	$2,142
Constrained WLS	0	2	-$0	-$1,538	-$0	-$378	-$0	-$8,888	$2,548	$1,467
Stacked LightGBM, constrained	0	2	$0	-$1,639	$0	-$583	-$0	-$9,294	$1,202	$317
Constrained WLS	0	3	$0	$803	-$0	$199	-$0	$1,853	$3,393	$4,150
Stacked LightGBM, constrained	0	3	$0	$592	-$0	-$140	-$0	$2,752	$2,293	$3,093
Constrained WLS	0	4	-$0	$51	-$0	$234	-$0	$1,145	$3,460	$6,334
Stacked LightGBM, constrained	0	4	$0	$273	$0	$632	-$0	$940	$2,097	$4,538
Constrained WLS	1	0	-$0	-$1,237	-$0	-$1,637	-$0	-$12,261	$2,773	$2,972
Stacked LightGBM, constrained	1	0	-$0	-$1,214	$0	-$1,349	-$0	-$11,664	$1,361	$1,300
Constrained WLS	1	1	$0	$208	-$0	$681	-$0	$5,057	$3,944	$2,670
Stacked LightGBM, constrained	1	1	-$0	$54	-$0	$192	-$0	$3,623	$2,346	$1,301
Constrained WLS	1	2	-$0	$93	-$0	$211	$0	$259	$3,370	$3,530
Stacked LightGBM, constrained	1	2	$0	$515	-$0	$565	-$0	$6,534	$2,173	$2,971
Constrained WLS	1	3	$0	$99	-$0	-$1,111	-$0	$840	$3,575	$3,516
Stacked LightGBM, constrained	1	3	$0	$91	-$0	-$1,084	$0	$2,348	$2,313	$3,332
Constrained WLS	1	4	$0	$870	-$0	$2,123	-$0	$1,728	$3,129	$3,952
Stacked LightGBM, constrained	1	4	-$0	$728	-$0	$2,102	-$0	$1,654	$1,964	$2,482
Constrained WLS	2	0	-$0	-$1,576	-$0	-$383	-$0	-$13,043	$2,468	-$128
Stacked LightGBM, constrained	2	0	$0	-$1,277	-$0	$124	-$0	-$10,379	$1,292	-$902
Constrained WLS	2	1	$0	$678	-$0	$491	-$0	$3,586	$3,694	$5,296
Stacked LightGBM, constrained	2	1	$0	$742	-$0	$917	-$0	$4,804	$2,443	$4,203
Constrained WLS	2	2	$0	$959	-$0	-$241	$0	$4,952	$3,782	$3,844
Stacked LightGBM, constrained	2	2	-$0	$984	-$0	-$22	-$0	$3,572	$2,331	$2,866
Constrained WLS	2	3	-$0	-$237	-$0	-$1,444	-$0	-$1,903	$3,284	$3,547
Stacked LightGBM, constrained	2	3	-$0	-$441	-$0	-$1,769	$0	-$406	$1,998	$1,868
Constrained WLS	2	4	$0	$406	-$0	$1,862	-$0	$3,740	$3,631	$4,089
Stacked LightGBM, constrained	2	4	-$0	$122	$0	$985	-$0	$1,953	$2,397	$2,748

In the training data the constraint holds exactly for the target groups. Out of fold it does not; the pooled values in Table 5 average across these folds.

Table 6: The 20 largest drivers of the squared-error LightGBM predictions

Rank	*Feature*	*Mean*	*SHAP*	
1	log_spend_y1	$4,390	$338	0.1083
2	log_rx_fills_y1	$1,762	$185	0.0101
3	n_conditions	$899	$194	0.0209
4	log_office_visits_y1	$525	$52	0.0039
5	log_inpatient_y1	$460	$102	0.0163
6	n_body_systems	$372	$33	-0.0119
7	END002+END005 Diabetes mellitus without complication / Diabetes mellitus, Type 2	$233	$46	-0.0013
8	age_75-84	$199	$57	-0.0004
9	age_85+	$164	$21	0.0101
10	age_55-64	$153	$50	0.0091
11	age_65-74	$142	$29	0.0044
12	MUS003 Rheumatoid arthritis and related disease	$136	$53	-0.0111
13	END010 Disorders of lipid metabolism	$132	$70	0.0098
14	age_5-17	$126	$15	-0.0026
15	female	$117	$72	0.0000
16	region_1.0	$99	$40	-0.0001
17	age_25-34_x_female	$98	$25	0.0163
18	END001 Thyroid disorders	$92	$40	0.0057
19	log_er_visits_y1	$90	$42	0.0016
20	region_3.0	$81	$21	-0.0002

Weighted mean absolute SHAP value on held-out folds. Rank stability across folds: mean Spearman correlation 0.79 over all features and 0.76 over the top 30. The last column is the loss in held-out R^2 when the feature is permuted in the CANN.

Table 7: Coding sensitivity: rise in predicted spending per added condition code, 10% of people gain one code (standard deviation across folds)

Features	*Model*	*Prevalence*	*Chronic*	*Targeted*	*Own-targeted*
F2	CANN	$1,919 (sd $281)	$3,372 (sd $547)	$6,483 (sd $1,500)	$8,467 (sd $1,987)
F2	Constrained WLS	$2,107 (sd $195)	$2,921 (sd $150)	$16,166 (sd $1,510)	$17,469 (sd $1,232)
F2	LightGBM (Tweedie)	$2,036 (sd $89)	$2,907 (sd $120)	$4,781 (sd $553)	$5,510 (sd $497)
F2	LightGBM (squared error)	$2,155 (sd $94)	$3,448 (sd $482)	$7,241 (sd $1,578)	$8,135 (sd $1,575)
F2	Payment-form WLS (non-negative)	$2,379 (sd $157)	$4,327 (sd $175)	$10,812 (sd $791)	$11,045 (sd $533)
F2	Tweedie GLM	$1,763 (sd $239)	$2,900 (sd $100)	$5,214 (sd $555)	$7,018 (sd $651)
F2	Two-part	$1,182 (sd $34)	$1,515 (sd $27)	$1,740 (sd $85)	$1,994 (sd $67)
F2	WLS	$1,837 (sd $194)	$2,806 (sd $269)	$10,529 (sd $818)	$10,777 (sd $531)
F3	CANN	$409 (sd $128)	$977 (sd $194)	$2,925 (sd $236)	$5,204 (sd $2,276)
F3	Constrained WLS	$154 (sd $167)	$111 (sd $284)	$12,812 (sd $1,377)	$13,754 (sd $1,158)
F3	LightGBM (Tweedie)	$616 (sd $49)	$1,010 (sd $266)	$1,482 (sd $228)	$2,323 (sd $313)
F3	LightGBM (squared error)	$789 (sd $91)	$1,290 (sd $217)	$3,046 (sd $1,299)	$4,776 (sd $1,235)
F3	Payment-form WLS (non-negative)	$1,287 (sd $101)	$2,905 (sd $200)	$9,418 (sd $827)	$9,806 (sd $656)
F3	Tweedie GLM	$397 (sd $112)	$953 (sd $178)	$2,916 (sd $234)	$5,055 (sd $2,083)
F3	Two-part	$827 (sd $128)	$1,283 (sd $338)	$2,071 (sd $759)	$2,572 (sd $944)
F3	WLS	$283 (sd $170)	$580 (sd $286)	$8,539 (sd $803)	$9,044 (sd $466)

Codes the person did not have are switched on; use and spending are unchanged. Prevalence: drawn in proportion to prevalence. Chronic: from payment-relevant chronic categories. Targeted: the five flags with the largest payment-form coefficients. Own-targeted: the five flags that raise that model's own prediction most, found on its training data. F2 excludes prior use and spending; F3 includes them.

Table 7b: Use sensitivity: next-year payment per added dollar of year-1 spending

Features	*Model*	*$ per $ of year-1 spending*	*SD across folds*	*Total rise*
F3	CANN	0.239	0.112	0.47%
F3	Constrained WLS	0.340	0.081	0.72%
F3	LightGBM (Tweedie)	0.342	0.071	0.71%
F3	LightGBM (squared error)	0.319	0.068	0.66%
F3	Payment-form WLS (non-negative)	0.010	0.004	0.02%
F3	Tweedie GLM	0.238	0.112	0.47%
F3	Two-part	0.039	0.017	0.08%
F3	WLS	0.280	0.068	0.57%

Year-1 spending raised by 20% for 10% of held-out people, nothing else changed. Only feature sets with prior spending respond.

boosting and WLS is −0.002 (−0.006 to 0.003). Adding social-risk and functional-status features (F4) changes R^2 by less than 0.01 for the linear and boosted models.

The Tweedie GLM collapses on F2, to an R^2 of −0.001 with a repeat standard deviation of 0.021 and a top-decile predictive ratio of 1.24. With a log link, the effects of co-occurring condition flags multiply, and without use terms to anchor the scale a few people with many rare flags receive very large predictions. The CANN and boosting, which start from or shrink toward simpler fits, do not show this. The payment-form model avoids negative predictions entirely, and its R^2 cost relative to WLS is small on F1 and F2 and larger on F3, where the unconstrained fit offsets overlapping use and condition terms with negative coefficients.

Payment formulas are calibrated on their age-sex cells, and Table A2b checks those cells. WLS, whose age-sex dummies force the cell means to balance in training, has predictive ratios within 0.01 of 1 in every cell out of fold. Tweedie LightGBM ranges from 0.92 (men 65 and over) to 1.13 (men 35 to 44), and the Tweedie GLM and CANN from 0.87 to 1.08. The payment-form model pays about 40% above cost for children of both sexes, because 42 of its 124 condition coefficients and half of its age cells sit at the zero bound, so young people are priced by use and count terms estimated mostly on adults; its predictions exceed total spending by 5%, which a payment system would remove with a normalization factor.

Table 7c: Own-targeted pools: the flags that raise each model's prediction most

Features	*Model*	*Flag (folds chosen, mean gain)*
F2	CANN	NVS009 (4, $9,146); CIR019 (3, $10,095); NEO030 (3, $8,924); MBD014 (3, $7,634); MUS003 (3, $6,524)
F2	Constrained WLS	NVS009 (5, $24,120); NVS015 (5, $18,247); MUS003 (5, $15,885); CIR019 (4, $15,448); NVS012 (3, $13,265)
F2	LightGBM (Tweedie)	NVS009 (5, $6,322); MUS003 (4, $6,046); NEO030 (4, $5,612); GEN006 (4, $5,171); RSP008 (2, $5,639)
F2	LightGBM (squared error)	NVS009 (5, $7,642); MUS003 (4, $9,373); GEN006 (4, $9,040); MUS026 (3, $6,716); CIR019 (2, $9,453)
F2	Payment-form WLS (non-negative)	MUS003 (5, $11,210); NVS009 (5, $10,512); NVS015 (4, $13,586); CIR019 (4, $10,873); GEN006 (4, $10,694)
F2	Tweedie GLM	NVS009 (5, $8,677); MBD014 (5, $7,717); MUS003 (4, $6,318); PRG029 (4, $6,054); NEO039 (3, $5,361)
F2	Two-part	NVS009 (5, $2,169); MUS026 (4, $2,071); MBD003 (4, $1,921); GEN006 (3, $2,152); MUS013 (2, $1,957)
F2	WLS	MUS003 (5, $11,183); NVS015 (4, $13,110); CIR019 (4, $10,745); GEN006 (4, $10,517); NVS009 (4, $10,016)
F3	CANN	NVS009 (5, $4,730); MUS003 (4, $3,806); MUS026 (4, $3,733); DEN001 (3, $14,319); GEN006 (3, $2,865)
F3	Constrained WLS	NVS009 (5, $17,944); NVS015 (5, $15,658); MUS003 (5, $12,890); CIR019 (4, $11,894); MUS026 (4, $10,754)
F3	LightGBM (Tweedie)	PRG029 (5, $3,017); MUS003 (4, $2,304); NVS009 (4, $2,149); MUS026 (4, $2,127); NVS017 (3, $2,058)
F3	LightGBM (squared error)	MUS003 (4, $5,025); NVS017 (4, $4,821); GEN006 (4, $4,593); PRG029 (4, $4,519); MUS026 (4, $4,506)
F3	Payment-form WLS (non-negative)	NVS015 (5, $11,826); MUS003 (5, $10,177); GEN006 (4, $9,928); CIR019 (3, $9,300); NVS009 (3, $8,977)
F3	Tweedie GLM	NVS009 (5, $4,627); MUS003 (4, $3,806); MUS026 (4, $3,760); GEN006 (4, $2,915); DEN001 (3, $13,678)
F3	Two-part	NVS009 (5, $3,049); MUS026 (4, $2,706); NEO039 (3, $2,774); GEN006 (3, $2,737); MUS003 (3, $2,562)
F3	WLS	MUS003 (5, $9,347); NVS015 (4, $11,901); GEN006 (4, $9,026); NVS009 (3, $7,743); CIR019 (2, $9,898)

For each fold, the five flags with the largest mean rise in that model's prediction when switched on for 3,000 training persons. Listed are the five chosen most often, with the number of folds (of five) and the mean training gain per code.

Who is underpaid (RQ2)

Table 4 and Figure 3 report net compensation by group, and Table 4b the paired difference from WLS. One result dominates. People who need help with ADL or IADL, with a mean annual cost of $31,329, are underpaid by every model: $7,747 per person-year by WLS (interval $4,246 to $12,747), $9,293 by the payment-form model, $4,762 by the Tweedie GLM, $5,721 by Tweedie LightGBM and $4,545 by the CANN, and between $4,545 and $9,583 across all ten. On the same resamples, the Tweedie GLM pays this group $2,985 more than WLS (interval $2,227 to $3,834) and Tweedie LightGBM $2,025 more ($1,354 to $2,925), while the payment-form model pays it $1,546 less. Flexible models narrow the gap without closing it, because functional status is not in the primary features and no combination of diagnoses and use reproduces it.

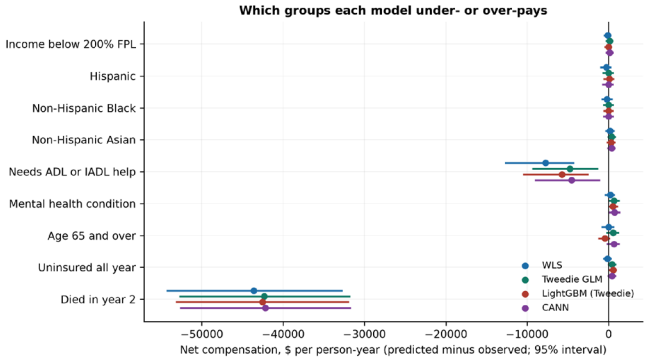


Figure 3: Net compensation by group and model, $ per person-year, with 95% intervals.

People who died in year 2 are underpaid by $42,174 to $46,604 per person-year of exposure by every model, predictive ratios of 0.30 to 0.36. No prospective model sees a death coming from the year before, and payment systems handle this through annualization and risk sharing rather than through the formula. The group is reported and not targeted.

The other groups are close to balance under most models, with intervals that include zero. WLS underpays Hispanic and non-Hispanic Black enrollees by $271 and $202, with intervals that include zero, and Tweedie LightGBM is within $91 of balance for both. Four patterns are outside noise. Tweedie LightGBM overpays people uninsured all year by $549 (interval $40 to $1,024), a predictive ratio of 1.23 on a mean cost of $2,401, because the low use of the uninsured is not fully carried into their next year. The random forest and CANN overpay people with mental health conditions by $971 and $724. The squared-error LightGBM underpays enrollees aged 65 and over by $1,004 ($336 to $1,832), and the random forest by $1,079. The payment-form model pays most groups above cost, by $413 to $780 for the race and ethnicity groups and $675 for low income, which is its 5% overall excess, about $400 per person, spread across groups.

The price of fairness (RQ3)

Table 5 and Figure 4 trace the frontier. For the linear formula, the largest target-group gap is the ADL gap, $7,747 at $\lambda = 0$. Penalizing it to $1,574 at $\lambda = 1$ reduces R^2 from 0.123 to 0.107, and the constrained limit leaves $724 at an R^2 of 0.099, a loss of a fifth of the formula's R^2. The stacked estimator starts from a gap of $4,601 at an R^2 of 0.182, reaches $536 at $\lambda = 1$ with an R^2 of 0.175, and $296 at 0.171 when constrained, a loss of 6% of its own R^2 and of 0.016 relative to the unconstrained boosted model. At every level of the gap the stacked frontier lies above the linear one. Decile recalibration of the boosted model, which uses no group information, keeps R^2 at 0.186 and leaves the gap at $5,608.

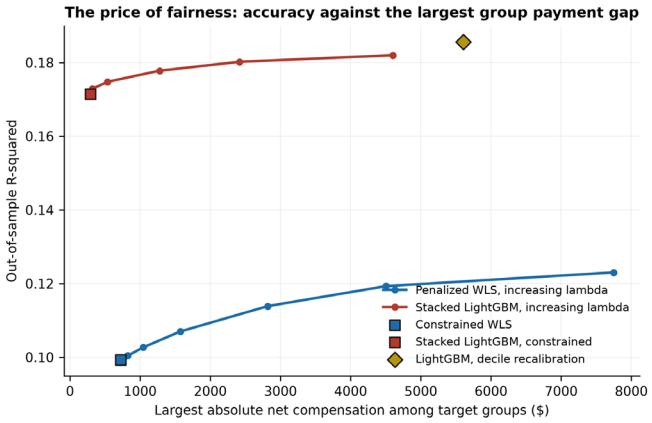


Figure 4: The accuracy-fairness frontier: out-of-sample R^2 against the largest absolute net compensation among target groups, penalized and constrained linear and stacked estimators, with 95% intervals.

The intervals for the gap are wide: $4,246 to $12,747 for WLS at $\lambda = 0$ and $220 to $5,515 under the constraint. For the constrained stacked estimator the interval, $365 to $4,590, lies above the point estimate of $296. That is a property of a maximum of absolute values near zero: resampling noise tends to push a maximum of absolute values up, so the upper limit is the informative bound, and it says the out-of-sample gap could be as large as $4,590.

Table 5b shows why. In the training data of every fold the constraint holds exactly. Out of fold, the ADL group's net compensation under constrained WLS ranges from −$13,043 to +$5,057 across the 15 folds, and under the constrained stacked estimator from −$11,664 to +$6,534. The pooled figure near zero is an average of large errors in both directions. With 1,612 persons in the group, about 320 in a held-out fold, a few very high spenders decide the fold's result. A constraint on a small high-variance group guarantees balance on the data used to fit it; it does not guarantee balance next year.

Closing the target gaps also opens others. Under constrained WLS, enrollees aged 65 and over move from balance to an overpayment of $3,362 (predictive ratio 1.20),

Table 8: Robustness: each row changes one decision

Variant	*Model*	*n*	*R^2*	*CPM*	*PR top decile*	*Max*
baseline	WLS	43,963	0.123	0.147	0.94	$7,830
baseline	Payment-form WLS (non-negative)	43,963	0.107	0.180	1.01	$9,375
baseline	Tweedie GLM	43,963	0.122	0.241	1.01	$4,809
baseline	LightGBM (Tweedie)	43,963	0.190	0.276	0.99	$5,886
outcome capped	WLS	43,963	0.192	0.173	0.93	$6,773
outcome capped	Payment-form WLS (non-negative)	43,963	0.173	0.195	0.99	$8,036
outcome capped	Tweedie GLM	43,963	0.180	0.247	1.00	$4,114
outcome capped	LightGBM (Tweedie)	43,963	0.281	0.284	1.00	$4,670
log outcome	WLS, log outcome	43,963	-0.096	0.193	1.63	$6,684
unweighted training	WLS	43,963	0.119	0.121	1.02	$6,288
unweighted training	Payment-form WLS (non-negative)	43,963	0.103	0.167	1.10	$7,963
unweighted training	Tweedie GLM	43,963	0.138	0.243	1.01	$3,489
unweighted training	LightGBM (Tweedie)	43,963	0.188	0.272	1.03	$3,919
CCSR floor 0.2%	WLS	43,963	0.125	0.153	1.02	$7,293
CCSR floor 0.2%	Payment-form WLS (non-negative)	43,963	0.115	0.190	1.05	$8,195
CCSR floor 0.2%	Tweedie GLM	43,963	0.123	0.245	1.02	$4,232
CCSR floor 0.2%	LightGBM (Tweedie)	43,963	0.191	0.278	0.98	$5,632
CCSR floor 1%	WLS	43,963	0.124	0.149	0.92	$8,510
CCSR floor 1%	Payment-form WLS (non-negative)	43,963	0.108	0.180	0.99	$10,013
CCSR floor 1%	Tweedie GLM	43,963	0.155	0.250	0.97	$5,631
CCSR floor 1%	LightGBM (Tweedie)	43,963	0.189	0.276	0.99	$5,857
no prior spending (F2)	WLS	43,963	0.103	0.187	1.05	$8,876
no prior spending (F2)	Payment-form WLS (non-negative)	43,963	0.098	0.191	1.05	$9,988
no prior spending (F2)	Tweedie GLM	43,963	0.011	0.176	1.24	$5,181
no prior spending (F2)	LightGBM (Tweedie)	43,963	0.101	0.199	1.00	$10,991
social-risk features (F4)	WLS	43,963	0.127	0.151	0.96	$1,165
social-risk features (F4)	Payment-form WLS (non-negative)	43,963	0.113	0.182	0.97	$3,244
social-risk features (F4)	Tweedie GLM	43,963	0.118	0.243	1.02	$1,426
social-risk features (F4)	LightGBM (Tweedie)	43,963	0.192	0.277	1.00	$2,237
decedents excluded	WLS	43,568	0.124	0.149	0.94	$7,887
decedents excluded	Payment-form WLS (non-negative)	43,568	0.106	0.180	1.01	$9,376
decedents excluded	Tweedie GLM	43,568	0.125	0.243	1.00	$5,015
decedents excluded	LightGBM (Tweedie)	43,568	0.193	0.277	0.98	$5,602
decedents not annualized	WLS	43,963	0.127	0.150	0.94	$7,422
decedents not annualized	Payment-form WLS (non-negative)	43,963	0.110	0.182	1.02	$8,909
decedents not annualized	Tweedie GLM	43,963	0.128	0.244	1.00	$4,554
decedents not annualized	LightGBM (Tweedie)	43,963	0.196	0.279	0.98	$5,359
3 folds	WLS	43,963	0.122	0.145	0.96	$7,642
3 folds	Payment-form WLS (non-negative)	43,963	0.107	0.181	1.01	$9,260
3 folds	Tweedie GLM	43,963	0.128	0.243	1.02	$3,785
3 folds	LightGBM (Tweedie)	43,963	0.187	0.278	0.99	$5,520
10 folds	WLS	43,963	0.125	0.150	0.94	$7,792
10 folds	Payment-form WLS (non-negative)	43,963	0.109	0.182	1.00	$9,311
10 folds	Tweedie GLM	43,963	0.115	0.242	0.99	$4,710

Variant	*Model*	*n*	R^2	*CPM*	*PR top decile*	*Max*
10 folds	LightGBM (Tweedie)	43,963	0.188	0.277	0.99	$5,430
panel 23 only	WLS	13,765	0.144	0.166	1.01	$9,996
panel 23 only	LightGBM (Tweedie)	13,765	0.220	0.290	1.07	$6,847
panel 24 only	WLS	9,568	0.070	0.086	1.16	$6,673
panel 24 only	LightGBM (Tweedie)	9,568	0.142	0.256	1.02	$4,432
panel 25 only	WLS	5,942	0.030	0.006	1.34	$2,999
panel 25 only	LightGBM (Tweedie)	5,942	0.155	0.230	1.00	$4,238
panel 26 only	WLS	6,579	0.032	0.012	1.15	$15,978
panel 26 only	LightGBM (Tweedie)	6,579	0.145	0.255	0.85	$16,961
panel 27 only	WLS	8,109	0.136	0.122	1.18	$3,516
panel 27 only	LightGBM (Tweedie)	8,109	0.248	0.287	0.97	$3,498
pandemic outcomes out	WLS	28,453	0.125	0.146	0.97	$9,285
pandemic outcomes out	Payment-form WLS (non-negative)	28,453	0.112	0.188	1.04	$10,300
pandemic outcomes out	Tweedie GLM	28,453	0.091	0.246	1.06	$6,060
pandemic outcomes out	LightGBM (Tweedie)	28,453	0.207	0.287	1.00	$7,697
age 65 and over	WLS	9,490	0.096	0.075	1.05	$6,011
age 65 and over	Payment-form WLS (non-negative)	9,490	0.072	0.088	1.04	$9,121
age 65 and over	Tweedie GLM	9,490	0.124	0.161	1.02	$4,463
age 65 and over	LightGBM (Tweedie)	9,490	0.156	0.185	0.93	$5,914
under 65, private	WLS	21,988	0.091	0.095	1.04	$7,811
under 65, private	Payment-form WLS (non-negative)	21,988	0.083	0.131	1.06	$8,783
under 65, private	Tweedie GLM	21,988	0.088	0.178	1.08	$3,535
under 65, private	LightGBM (Tweedie)	21,988	0.155	0.210	0.98	$6,156

One repeat of the survey folds. Max |NC| is over the three target groups.

Hispanic enrollees to $1,263 below cost, non-Hispanic Black enrollees to $565 below cost, and the uninsured to $1,498 below cost, a predictive ratio of 0.38. The stacked constraint moves the same groups by smaller amounts: $2,242 for the elderly, and −$685, −$290 and −$601 for Hispanic, non-Hispanic Black and uninsured enrollees. The ADL group is largely elderly, and a constraint that raises payment for people who need help raises payment for the elderly generally and takes it from young, low-cost groups. We report all groups at every point of the frontier because this spillover is invisible if only target groups are shown.

Drivers (RQ4)

Table 6 and Figure 5 list the largest contributors to the squared-error LightGBM. Year-1 spending is first by a wide margin (mean absolute SHAP $4,390), followed by prescription fills ($1,762), the condition count ($899), office visits ($525), inpatient discharges ($460) and the body-system count ($372). The first diagnosis category is diabetes ($233), then the age bands from 55 upward. Importance ranks are stable across folds (Spearman 0.79 over all features, 0.76 over the top 30). SHAP contributions rise with year-1 spending, condition count and body-system count (rank correlations 0.88, 0.85 and 0.86; Table A4). For the CANN, permuting year-1 spending costs 0.108 of R^2 and no other feature costs more than 0.021.

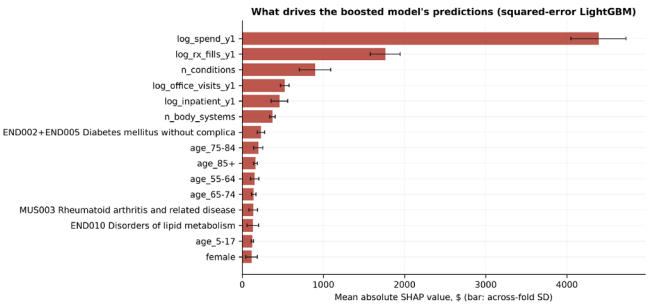


Figure 5: The fifteen largest drivers of the squared-error LightGBM predictions, mean absolute SHAP value in dollars with across-fold standard deviation.

The drivers explain the accuracy results. What the boosted model knows beyond a linear formula is how prior spending and use carry forward, nonlinearly and with interactions. On F2, without those inputs, it has no advantage, and the diagnosis flags it relies on are the same ones a linear formula uses.

Use sensitivity (RQ5)

Table 7b gives the rise in next-year predicted spending per added dollar of year-1 spending on F3. Tweedie LightGBM pays 34 cents (fold standard deviation 7), constrained WLS 34 cents, squared-error LightGBM 32, WLS 28, and the Tweedie GLM and CANN 24. The payment-form model pays 1 cent and the two-part model 4. These last two are not immune to use: the payment-form model loads prior use on inpatient discharges, prescription fills and office visits, which the experiment holds fixed, and puts almost no weight on logged spending itself. A payer that raises its members' spending by a dollar recovers a third of it next year from a boosted model on F3, before any rise in visits and fills. On F2 no model responds, by construction.

Coding sensitivity (RQ5)

Table 7 and Figure 6 give the rise in predicted spending per added code when 10% of held-out people gain one, Table A3 the full grid, and Table 7c the codes each model pays most. The dollars per code change little with the share of people affected or with two codes instead of one.

Codes drawn at random

With prior use in the model (F3), a code drawn in proportion to prevalence raises predicted spending by $283 per code under WLS, $616 under Tweedie LightGBM and $1,287 under the payment-form model, and a chronic code by $580, $1,010 and $2,905. Unconstrained WLS responds least: it prices next-year cost mainly through prior spending and the condition count, with coefficients that offset each other, so a flag added to someone whose recorded use has not changed moves it little. The payment-form model responds most, because none of its coefficients may be negative and so none can offset. Without prior use (F2) the models converge: a chronic code raises WLS by $2,806, the Tweedie GLM by $2,900, Tweedie LightGBM by $2,907 and the payment-form model by $4,327.

Codes chosen from a published formula

The five categories with the largest payment-form coefficients are polyneuropathies, rheumatoid arthritis, other diseases of the kidney and ureters, epilepsy and heart failure, each held by 0.6% to 1.7% of the sample. Adding one of those to a person who lacks it raises predictions on F3 by $8,539 under WLS, $9,418 under the payment-form model and $12,812 under constrained WLS, against $1,482 under Tweedie LightGBM. On F2 the same codes are worth $10,529, $10,812 and $16,166 against $4,781.

Codes chosen for each model

That pool is the payment-form model's own, which favors the other models. When each model is gamed instead on the five flags that raise its own prediction most, found on 3,000 of its own training persons (Table 7c), the ordering holds. On F3 an own-targeted code is worth $2,323 under Tweedie LightGBM, $2,572 under the two-part model, $4,776 under squared-error LightGBM, $5,055 under the Tweedie GLM, $5,204 under the CANN, $9,044 under WLS, $9,806 under the payment-form model and $13,754 under constrained WLS. On F2 the figures are $5,510 for Tweedie LightGBM, $7,018 for the Tweedie GLM, $10,777 for WLS, $11,045 for the payment-form model and $17,469 for constrained WLS. When 10% of people each gain one such code on F2, total predicted spending rises by 7.0% under Tweedie LightGBM, 13.7% under WLS, 14.1% under the payment-form model and 25.0% under constrained WLS. The linear models converge on the same five codes in almost every fold; the boosted and neural models' lucrative codes vary more across folds and include categories such as muscle disorders, nerve root disorders and uncomplicated pregnancy, which is why their fold standard deviations are larger.

Why

A linear formula attaches a fixed increment to each flag, and a rare, costly category carries a large increment estimated from a few hundred people; anyone coded with it receives the full amount. A tree splits on a rare flag only where enough people share it, and it conditions on use, so a code with no matching visits, fills or spending moves it less. The same property that makes the boosted model accurate on F3 makes it hard to move with a code alone. Constrained WLS is the most exposed of all, and the reason is the constraint: epilepsy, polyneuropathies, heart failure and rheumatoid arthritis are 6 to 11 times more common among people who need help with daily activities than among everyone else, so a constraint that raises payment for that group raises exactly the coefficients a coder would target. Fairness bought through a linear formula's diagnosis coefficients is bought with exposure to those diagnoses.

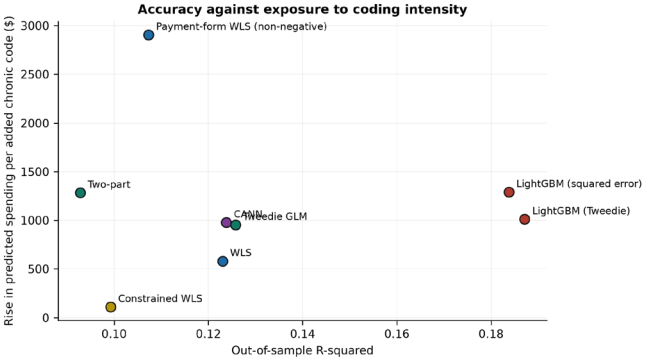


Figure 6: Accuracy against exposure to coding intensity: out-of-sample R^2 and the rise in predicted spending per added code when 10% of people gain one.

Robustness

Table 8 reports the variants. Across every variant except F2 the boosting advantage over WLS on R^2 is between 0.060 and 0.125, and on F2 it vanishes. Capping the outcome at $250,000 raises every R^2 (Tweedie LightGBM to 0.281, WLS to 0.192) and reduces the ADL gap by $700 to $1,200 for WLS,

the Tweedie GLM and boosting. Dropping prior use and spending (F2) raises the ADL gap for boosting to $10,991, since prior use was carrying part of the functional-status signal. Adding social-risk and functional-status features (F4), which include the ADL and IADL items, cuts the ADL gap to $1,165 for WLS and $2,237 for boosting: the direct route to paying for functional need is to observe it. Dropping decedents (Tweedie LightGBM 0.193, WLS 0.124) or not annualizing their spending (0.196 and 0.127) moves R^2 by less than 0.01. Three or ten folds change R^2 by less than 0.01. Unweighted training and moving the condition-flag floor change R^2 by less than 0.02 for WLS and boosting, but the Tweedie GLM rises from 0.122 to 0.155 with a 1% floor, another sign of its sensitivity to rare flags. The log-outcome linear model performs poorly on squared error (R^2 −0.096) and is not a candidate for payment.

Single panels vary more than any modeling choice: WLS ranges from 0.030 to 0.144 and Tweedie LightGBM from 0.142 to 0.248 across the five panels, and the ADL gap from about $3,000 to $17,000. Dropping outcome years 2020 and 2021 gives Tweedie LightGBM 0.207 and WLS 0.125. Within the 65-and-over subset, R^2 falls for every model (Tweedie LightGBM 0.156, Tweedie GLM 0.124, WLS 0.096) and the ADL gap persists; among people under 65 with private coverage, Tweedie LightGBM reaches 0.155 and WLS 0.091.

Discussion

For payers and regulators

On this benchmark the choice between a linear formula and a flexible model is not a choice between accuracy and fairness. The flexible model is more accurate when it uses prior use, it is no worse and for the functional-need group somewhat better on group compensation, and its drivers are stable and sensible. But its accuracy advantage comes entirely from prior use, and with prior use comes a payment that rises by about a third of a dollar for every dollar of this year's spending. A regulator who keeps the exclusion of prior use should expect the flexible model's accuracy advantage to vanish (Table 3). On coding, the ordering depends on which codes are added. Codes drawn at random move every model by similar amounts once prior use is excluded, but codes chosen for what they pay move a linear formula two to four times as far as a boosted model, and a fairness-constrained linear formula furthest of all. A regulator who constrains a formula for fairness buys the gap reduction with exposure to the very diagnoses that close it, and should audit those codes.

Three findings apply to any formula. First, the group every model underpays is defined by functional status, which no diagnosis or use variable reproduces. Observing it directly (F4) removes most of the gap without a constraint. Second, fairness constraints are fragile out of sample for small, costly groups. A regulator who imposes a constraint should expect next year's net compensation for the group to vary by several thousand dollars around zero, and should pair a constraint with risk sharing for the extreme cases it cannot reach. Third, fairness reporting must cover non-target groups with intervals. The constraint that fixed the ADL gap created an elderly overpayment of $3,362 and a 62% underpayment of the uninsured, and a reader shown only the target groups would call the constrained formula fair.

Questions a reader will ask

Are these R^2 values low? They are in the range expected for prospective prediction of untruncated annual spending from demographics and diagnoses (Ellis, Martins and Rose, 2018). MEPS conditions are self-reported and coded to three digits, so the levels are not comparable with claims-based studies; comparisons are valid within the benchmark, where every model sees the same data, folds and resamples.

Is the boosting gain real or noise? The paired interval for the gain over WLS on F3 excludes zero and no resample favors WLS, and the gain holds across fold counts, weighting, outcome caps and panels. The equally clear result is that there is no gain on F2.

Would a payer use unconstrained WLS? No. It predicts negative payments for 15.2% of people. It is reported because cost-prediction studies usually compare against it, and the payment-form model is the comparison a regulator should read.

Why keep decedents? Because payment models keep partial-year enrollees and annualize their spending. Excluding them or not annualizing changes R^2 by less than 0.01 and the ADL gap by a few hundred dollars.

Is the fairness frontier an artifact of the chosen targets? The targets were chosen before the analysis, and all nine groups are reported at every point. Different targets would move other groups; the spillover to the elderly and the uninsured follows from how the ADL group is composed and would appear under any constraint that pays more for functional need through age and diagnosis terms.

Is the targeted coding pool rigged against the linear formula? It would be if the only targeted pool were the payment-form model's own largest coefficients, so each model is also gamed on the five codes that raise its own prediction most, found on its own training data. The ordering is the same, and the gap between the linear forms and boosting narrows but does not close.

Why is the stacked estimator's gap interval above its point estimate? Because the statistic is a maximum of absolute values near zero, whose bootstrap distribution is pushed upward. The per-fold table is the more direct evidence of instability.

Limitations

MEPS is a household survey with self-reported conditions verified in part by provider records, and its three-digit ICD-10 codes make condition categories coarser than payment-model categories. Sample sizes for small groups are limited: the ADL group has 1,612 persons and the decedents 395, and their intervals are wide. The bootstrap holds fitted models

fixed; the repeat spread adds refitting variation, which is small except for the neural network. The fairness targets were chosen by the analyst. The coding experiment switches codes on mechanically: it measures what a formula would pay for added codes, not how coders would respond, and it does not model the documentation and audits that a real coder faces. The own-targeted pools are found by querying each fitted model, which a coder outside the payer could not do directly, though the behavior it stands for, finding what pays and recording it, needs no such access. The use experiment raises spending without raising visits and fills, so it understates the response of models that load on counts. The neural models were tuned over a small grid, and their results are a floor on what a well-tuned network can do. The analysis does not model plan selection or enrollee choice; a model's net compensation for a group is what gives plans an incentive, not a measure of how plans respond.

What the toolkit is for

Any payer or regulator with person-level cost data can run the same five evaluations on its own formula and candidate replacements, with survey- or cluster-aware validation, paired intervals and full group reporting built in. The package does not decide what fairness means; it measures what a formula does to each group, whether that holds out of sample, what changing it would cost, and how the formula responds to coding and to use.

Conclusion

A public benchmark shows that flexible models predict next-year health spending more accurately than the linear form of payment formulas only when they use prior utilization, that every model underpays people who need functional help unless it observes functional status, that fairness constraints close target gaps in training but not reliably out of sample and shift payment toward the elderly and away from the uninsured, and that a linear payment form pays two to four times as much as a boosted model for codes chosen to game it, with the fairness-constrained formula the most exposed of all. The code that produced every number is released as a reusable package.

Declarations

Data availability. MEPS public-use files are freely available from the Agency for Healthcare Research and Quality; the CCSR reference file from AHRQ's HCUP. Neither is redistributed.

Code availability. The `riskfair` package and the complete, seeded pipeline are at https://github.com/tosin-babs/fair-ml-health-risk-adjustment; `python/run_all.py` rebuilds every table and figure and `pytest tests` runs the checks. An interactive explorer of the results is at https://fair-risk-adjustment.vercel.app. It is a research benchmark on survey data, not a payment formula.

Competing interests. None declared.

Ethics. The analysis uses de-identified public survey data and did not require ethical approval.

References

[1] Andriola, C., Ellis, R. P., Siracuse, J. J., Hoagland, A., et al. (2024). A novel machine learning algorithm for creating risk-adjusted payment formulas. *JAMA Health Forum*, 5(4), e240625. doi:10.1001/jamahealthforum.2024.0625

[2] Bergquist, S. L., Layton, T. J., McGuire, T. G., & Rose, S. (2019). Data transformations to improve the performance of health plan payment methods. *Journal of Health Economics*, 66, 195–207. doi:10.1016/j.jhealeco.2019.05.005

[3] Brown, J., Duggan, M., Kuziemko, I., & Woolston, W. (2014). How does risk selection respond to risk adjustment? New evidence from the Medicare Advantage program. *American Economic Review*, 104(10), 3335–3364. doi:10.1257/aer.104.10.3335

[4] Centers for Medicare & Medicaid Services (2023). *Announcement of Calendar Year (CY) 2024 Medicare Advantage (MA) Capitation Rates and Part C and Part D Payment Policies*. 31 March 2023. Baltimore, MD: CMS.

[5] Cumming, R. B., Knutson, D., Cameron, B. A., & Derrick, B. (2002). *A Comparative Analysis of Claims-Based Methods of Health Risk Assessment for Commercial Populations*. Schaumburg, IL: Society of Actuaries. https://www.soa.org/globalassets/assets/files/research/projects/risk-assessmentc.pdf

[6] Duncan, I., Loginov, M., & Ludkovski, M. (2016). Testing alternative regression frameworks for predictive modeling of health care costs. *North American Actuarial Journal*, 20(1), 65–87. doi:10.1080/10920277.2015.1110491

[7] Ellis, R. P., Martins, B., & Rose, S. (2018). Risk adjustment for health plan payment. In T. G. McGuire & R. C. van Kleef (Eds.), *Risk Adjustment, Risk Sharing and Premium Regulation in Health Insurance Markets* (pp. 55–104). Academic Press. doi:10.1016/B978-0-12-811325-7.00003-8

[8] Ellis, R. P., Martins, B., & Zhu, W. (2017). Health care demand elasticities by type of service. *Journal of Health Economics*, 55, 232–243. doi:10.1016/j.jhealeco.2017.07.007

[9] Ellis, R. P., & McGuire, T. G. (1986). Provider behavior under prospective reimbursement: cost sharing and supply. *Journal of Health Economics*, 5(2), 129–151. doi:10.1016/0167-6296(86)90002-0

[10] Ellis, R. P., & McGuire, T. G. (1993). Supply-side and demand-side cost sharing in health care. *Journal of Economic Perspectives*, 7(4), 135–151. doi:10.1257/jep.7.4.135

[11] Geruso, M., & Layton, T. (2020). Upcoding: evidence from Medicare on squishy risk adjustment. *Journal of Political Economy*, 128(3), 984–1026. doi:10.1086/704756

[12] Irvin, J. A., Kondrich, A. A., Ko, M., Rajpurkar, P., et al. (2020). Incorporating machine learning and social determinants of health indicators into prospective risk adjustment for health plan payments. *BMC Public Health*, 20, 608. doi:10.1186/s12889-020-08735-0

[13] Jacobs, P. D., & Kronick, R. (2018). Getting what we pay for: how do risk-based payments to Medicare Advantage plans compare with alternative measures of beneficiary health risk? *Health Services Research*, 53(6), 4997–5015. doi:10.1111/1475-6773.12977

[14] Kautter, J., Pope, G. C., Ingber, M., Freeman, S., Patterson, L., Cohen, M., & Keenan, P. (2014). The HHS-HCC risk adjustment model for individual and small group markets under the

Affordable Care Act. *Medicare & Medicaid Research Review*, 4(3), E1–E46. doi:10.5600/mmrr.004.03.a03

[15] Kronick, R., & Welch, W. P. (2014). Measuring coding intensity in the Medicare Advantage program. *Medicare & Medicaid Research Review*, 4(2), E1–E19. doi:10.5600/mmrr.004.02.sa06

[16] Layton, T. J., Ellis, R. P., McGuire, T. G., & van Kleef, R. (2017). Measuring efficiency of health plan payment systems in managed competition health insurance markets. *Journal of Health Economics*, 56, 237–255. doi:10.1016/j.jhealeco.2017.05.004

[17] McGuire, T. G., Zink, A. L., & Rose, S. (2021). Improving the performance of risk adjustment systems: constrained regressions, reinsurance, and variable selection. *American Journal of Health Economics*, 7(4), 497–521. doi:10.1086/716199

[18] Medicare Payment Advisory Commission (2025). The Medicare Advantage program: status report. In *Report to the Congress: Medicare Payment Policy*, March 2025, chapter 11, pp. 317–406. Washington, DC: MedPAC.

[19] Montz, E., Layton, T., Busch, A. B., Ellis, R. P., Rose, S., & McGuire, T. G. (2016). Risk-adjustment simulation: plans may have incentives to distort mental health and substance use coverage. *Health Affairs*, 35(6), 1022–1028. doi:10.1377/hlthaff.2015.1668

[20] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. doi:10.1126/science.aax2342

[21] Park, S., & Basu, A. (2018). Alternative evaluation metrics for risk adjustment methods. *Health Economics*, 27(6), 984–1010. doi:10.1002/hec.3657

[22] Pope, G. C., Kautter, J., Ellis, R. P., et al. (2004). Risk adjustment of Medicare capitation payments using the CMS-HCC model. *Health Care Financing Review*, 25(4), 119–141.

[23] Rao, J. N. K., & Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401), 231–241. doi:10.1080/01621459.1988.10478591

[24] Rose, S. (2016). A machine learning framework for plan payment risk adjustment. *Health Services Research*, 51(6), 2358–2374. doi:10.1111/1475-6773.12464

[25] Schelldorfer, J., & Wüthrich, M. V. (2019). Nesting classical actuarial models into neural networks. SSRN working paper 3320525. doi:10.2139/ssrn.3320525

[26] Shrestha, A., Bergquist, S., Montz, E., & Rose, S. (2018). Mental health risk adjustment with clinical categories and machine learning. *Health Services Research*, 53(S1), 3189–3206. doi:10.1111/1475-6773.12818

[27] van Kleef, R. C., McGuire, T. G., van Vliet, R. C. J. A., & van de Ven, W. P. M. M. (2017). Improving risk equalization with constrained regression. *European Journal of Health Economics*, 18(9), 1137–1156. doi:10.1007/s10198-016-0859-1

[28] Wüthrich, M. V., & Merz, M. (2023). *Statistical Foundations of Actuarial Learning and its Applications*. Springer. doi:10.1007/978-3-031-12409-9

[29] Zink, A., & Rose, S. (2020). Fair regression for health care spending. *Biometrics*, 76(3), 973–982. doi:10.1111/biom.13206

[30] Zink, A., & Rose, S. (2021). Identifying undercompensated groups defined by multiple attributes in risk adjustment. *BMJ Health & Care Informatics*, 28(1), e100414. doi:10.1136/bmjhci-2021-100414

[31] Agency for Healthcare Research and Quality. *Medical Expenditure Panel Survey, Panel 23–27 Longitudinal Data Files (HC-217, HC-225, HC-234, HC-244, HC-252) and Medical Conditions Files (HC-207, HC-214, HC-222, HC-231, HC-241).*

[32] Agency for Healthcare Research and Quality, Healthcare Cost and Utilization Project. *Clinical Classifications Software Refined (CCSR) for ICD-10-CM Diagnoses, v2026.1*.