Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Neighborhood Health With Machine Learning and Flexible Regression Across U.S. Census Tracts

Shervin Assari

shervinassari@cdrewu.edu

Departments of Internal Medicine, Psychiatry, and Urban Public Health, Charles R. Drew University of Medicine and Science, Los Angeles, California, USA; Marginalization-Related Diminished Returns (MDRs) Center, Los Angeles, California, USA https://orcid.org/0000-0002-5054-6250

Research Article

Keywords: small-area estimation, neighborhood health, machine learning, geographic validation, socioeconomic conditions, census tract

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11827168/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Objective.

To determine which prediction methods generalize best to census tracts in held-out states, which health outcomes gain most from flexible modeling, and which socioeconomic and demographic predictors contribute most when modeled as curves.

Methods.

We linked ten socioeconomic and demographic measures from the 2014–2018 American Community Survey (ACS) to crude adult prevalence estimates from the 2025 CDC PLACES release. In five state-grouped folds, we compared linear regression, additive splines, splines plus pairwise products, extremely randomized trees, and histogram gradient boosting. We evaluated predictor-specific curvature by replacing one spline at a time with a straight line and refitting the model. Exploratory analyses compared linear, quadratic, and spline terms for tract age composition and assessed the incremental value of age interactions.

Results.

Across 15 outcomes, adding curves and pairwise interactions reduced root mean squared error (RMSE) by 7.83% on average compared with ordinary linear regression. Gradient boosting reduced error by 7.35%, additive splines by 6.03%, and extremely randomized trees by 5.13%. The largest gains were for diabetes (12.4%), stroke (12.0%), and chronic obstructive pulmonary disease (11.5%). Cancer had the highest held-out R^2 (0.930). Restoring a curved term for the percentage of Latino residents, median household income, or college attainment reduced error by about 1% on average for each predictor. For cancer, allowing both age percentages to curve reduced error by 4.10% relative to linear age terms.

Conclusions.

An interpretable regression with curves and pairwise interactions had the lowest held-out error for 10 of 15 outcomes. Prediction gains depended on the health measure and on which predictor was allowed to curve. The findings concern geographic variation in published prevalence estimates; independent health measurements are needed to evaluate accuracy for underlying disease burden.

Introduction

For public health planning, an accurate estimate of the distribution of illness across neighborhoods can matter even when individual records are unavailable. A straight-line model assumes that the same ten-point difference in college attainment, for example, carries the same predicted prevalence difference at

either end of the attainment distribution. A flexible model can relax that assumption. It may also learn that two area characteristics jointly predict prevalence in ways a simple sum cannot capture.

Machine learning (ML) has been used to estimate neighborhood obesity prevalence from features of the built environment visible in satellite images [1]. Related work combined satellite imagery and household surveys to estimate local economic conditions where direct measurements were scarce [2]. In both settings, the choice of inputs and the geographic scale were part of the prediction problem. With ten familiar census indicators rather than thousands of image features, a public health investigator faces a different choice: how much predictive information is lost when each indicator is assumed to have a constant slope? A fair test of ML in this setting requires an alternative that can already represent plausible curves.

CDC PLACES makes it possible to compare conditions across the same geographic units using a common small-area estimation framework [3, 4]. That consistency presents an opportunity as well as a methodological challenge. For instance, if a flexible model predicts diabetes particularly well from the age and income composition of a tract, some of its success may come from approximating relationships built into PLACES estimates. The relevant question is how accurately a method predicts the published tract estimate in a new state. Comparing its predictions against independent measures of disease would require a separate study.

The way a predictor enters a model matters. Dividing a continuous variable into groups can conceal variation within groups and make results depend on the chosen cutoffs [5]. Spline regression offers a way to estimate smooth changes in slope across the observed distribution [6]. Curvature and interaction answer different questions. An age-composition curve allows the difference associated with a ten-point higher percentage of older adults to depend on the starting percentage. An age-by-income interaction allows the difference to depend on tract income. Interaction is scale-dependent, and reporting the scale is essential to interpretation [7, 8]. Here the target is prevalence in percentage points and the criterion for retaining flexibility is improved prediction in held-out states.

In epidemiology, a prediction question calls for a different evaluation standard from a question about etiology [9]. Medical applications have used ML to estimate future risk [10] and extract predictive information from complex measurements [11]. Reviews of high-performance medicine also describe the importance of assessing the role an algorithm would play alongside existing practice [12]. For neighborhood surveillance the comparator is already a functioning statistical estimation system. Our narrower task is to predict its published estimates from earlier area characteristics and quantify the added value of model flexibility.

The distinction is important for interpreting prediction research. A condition can be highly predictable because its prevalence is closely related to age, yet gain little from machine learning if a line approximates that relationship well. Another condition can benefit more from flexible modeling while remaining difficult to predict in absolute terms. We therefore distinguished overall accuracy from

improvement beyond a linear baseline. The unit of prediction was a census tract in a state excluded from model fitting.

Successful ML applications do not by themselves establish that a tree ensemble will outperform regression on tract data. Deep learning has performed well in tightly defined image-based tasks such as detection of diabetic retinopathy [13] and skin-lesion classification [14]. In contrast, a systematic review of clinical prediction studies found little evidence of a consistent advantage over logistic regression in comparisons at lower risk of bias [15]. Neither type of evidence determines what to expect for a continuous, modeled neighborhood outcome. That comparison has to be made for the same places, outcomes, and validation design.

We organized the analysis around three aims. First, we compared two tree-based machine-learning methods with linear regression and with regression that permits curvature and interactions. Second, we identified which outcomes were most accurately predicted and which gained the most over the linear baseline. Third, for each outcome, we examined whether allowing a specific predictor to curve improved predictions after the other nine predictors were already modeled flexibly. All three comparisons used the same states, tracts, and folds.

Methods

Data and sample. We matched 60,168 historical ACS and 2025 PLACES tract identifiers and retained 56,112 complete tracts in 48 state-level jurisdictions. Of 83,522 PLACES records, 23,354 did not match the historical identifiers and 4,056 additional matched records lacked complete model data. Identifier matching does not guarantee unchanged boundaries. PLACES publishes small-area estimates based on survey and population data, including demographic and socioeconomic inputs [3, 4]. The outcomes were 15 published crude prevalences. Higher values indicate a greater modeled burden. This study predicts published estimates and cannot independently validate observed disease.

Predictors. We included median household income; percentages living at or above poverty, college educated (ages 25+), and civilian employed (ages 16+); percentages of non-Hispanic Black, non-Hispanic White and Hispanic residents; the percentage female; and percentages of adults aged 45–64 and 65 or older. Income was measured in 2018 dollars. All percentages used their corresponding ACS denominators. No other health outcome, state indicator, geographic coordinates, or PLACES prevalence was supplied as a test-tract predictor. ACS detailed tables were B01001, B03002, B15003, B17001, B19013 and B23025 [16].

Validation. A fixed shuffled five-fold split (seed 641) assigned entire states to folds. Every tract received one out-of-fold prediction. Transformations, knot values, and regression coefficients were learned on training tracts. The primary summary weights tracts equally within each outcome and outcomes equally in the overall comparison. As a geographic sensitivity analysis we also gave each state equal weight when averaging mean squared errors. Settings were fixed; performance comparisons describe the fitted

configurations. The best model for any outcome was identified after examining its held-out errors, making that ranking exploratory.

Holding out whole states addresses a practical concern about geographic prediction. With random tract-level splits, a training set can contain many tracts that share local institutions, labor markets, or patterns of survey estimation with the test tracts. Geographic blocking is one approach to evaluating transfer beyond the sampled locations [17]. Here the holdout was geographic rather than temporal. The results apply to prediction of the same PLACES release in the represented jurisdictions and do not establish forecasting accuracy for a future year.

Models. Ordinary least squares served as baseline. Additive cubic splines used up to six distinct training quantile knots per predictor and linear extrapolation. A second spline model added all 45 pairwise products of centered, standardized predictors. The multioutput extremely randomized tree ensemble used 60 trees, minimum leaf size 15 and 90% of features per split, standardizing targets on training data. Separate histogram gradient boosting models used 140 iterations, learning rate 0.08, up to 15 leaves, minimum leaf size 35, and L2 penalty 1. These settings were held constant across folds. Comparing the two spline models tests a specified class of interactions; trees can represent other interactions or thresholds.

The sequence of models makes the first aim interpretable. An additive model can describe diminishing or accelerating differences along a predictor without requiring the slope to depend on a second predictor [18]. Adding product terms allows, for example, the predicted income gradient to differ with age composition. Extremely randomized trees combine predictions from trees grown with randomized splits [19]. Gradient boosting fits trees sequentially to improve the current prediction [20]. These methods offer different routes to flexibility, so a comparison with both regression benchmarks is more informative than a comparison with a straight line alone.

Performance and uncertainty. RMSE and mean absolute error are in prevalence percentage points. Pooled held-out R^2 is relative to variation around the pooled observed mean. Relative improvement is 100 × (1 − RMSE of candidate / RMSE of comparator). For each model–outcome pair we resampled the 48 held-out state error totals with replacement 2,000 times. Percentile intervals describe variation across test states conditional on the fitted fold models. They do not include retraining uncertainty, ACS sampling error, uncertainty in the PLACES estimates, or adjustment for model selection.

Predictor analysis. For each of ten predictors we refitted the additive model with that predictor constrained to be linear, retaining flexible terms for the other nine. The RMSE reduction obtained by restoring the spline estimates its incremental predictive contribution from curvature. This is different from total predictor importance. Correlated measures can substitute for one another, and these comparisons cannot attribute interactions or identify causal effects.

Exploratory age and interaction analyses. We compared separate linear, quadratic, and cubic spline terms for the percentages of adults aged 45–64 and 65 or older, with the other eight predictors modeled

using splines. We next removed each of the 45 products from the full interaction model, and then removed blocks of products involving age. All main effects remained in the model. As in the primary analysis, a positive gain denotes the percentage reduction in held-out RMSE obtained by restoring the omitted terms. These comparisons ask what the terms add after the remaining model has been fitted; gains from overlapping blocks cannot be summed.

Results

Aim 1. Which approach generalized best across states?

Splines + products had the largest mean gain, with lower RMSE than linear regression for 15 of 15 outcomes. The second-highest average was boosting. Table 1 weights each outcome equally, preventing outcomes with greater variability from dominating the comparison.

Table 1
Five-fold held-out results. Positive values indicate lower error than linear regression; wins are descriptive rankings.

Model	Mean RMSE gain (%)	State-equal gain (%)	Lowest RMSE (outcomes)
Linear	0.00	0.00	0
Additive splines	6.03	5.94	0
Splines + products	7.83	8.74	10
Extra trees	5.13	6.20	2
Boosting	7.35	8.31	3

Giving states equal weight emphasized performance in less populous jurisdictions. The leading methods remained regression with curves and interactions (8.74% mean RMSE reduction) and boosting (8.31%). Under this weighting, extremely randomized trees (6.20%) slightly exceeded additive splines (5.94%). The method with the smallest average error therefore depended little on weighting, although the ordering of the other methods was less stable.

Aim 2. Which outcomes benefited most?

Diabetes, stroke, and COPD gained the most from relaxing the linear model. For diabetes, RMSE fell from 2.02 to 1.77 prevalence points, a 12.4% reduction (95% state-resampling interval, 9.7%–14.4%). For stroke it fell from 0.590 to 0.519 points, a 12.0% reduction (10.8%–13.2%). For COPD it fell from 1.49 to 1.32 points, an 11.5% reduction (9.7%–13.3%). Relative reductions make outcomes easier to compare; the absolute differences show what the improvement means on each prevalence scale.

Table 2
Each row uses the model with the lowest observed RMSE among the five compared. RMSE is in percentage points.

Outcome	Lowest-error model	RMSE	Gain (%)	R^2
Diabetes	Splines + products	1.77	12.4	0.807
Stroke	Splines + products	0.52	12.0	0.831
COPD	Splines + products	1.32	11.5	0.779
Cancer	Splines + products	0.71	10.8	0.930
Fair/poor health	Splines + products	3.20	10.3	0.825
Coronary heart disease	Splines + products	0.89	8.6	0.804
High blood pressure	Splines + products	3.17	8.1	0.812
Physical distress	Boosting	1.49	7.9	0.792
Arthritis	Splines + products	2.54	7.5	0.837
Any disability	Extra trees	3.73	7.1	0.765
Obesity	Splines + products	4.06	7.1	0.709
Current asthma	Splines + products	0.90	7.0	0.593
High cholesterol	Boosting	2.48	5.9	0.694
Mental distress	Boosting	1.54	2.5	0.765
Depression	Extra trees	3.11	2.5	0.427

Cancer was the most accurately predicted outcome by R^2: linear regression already explained 91.2% of between-tract variation, rising to 93.0% with curves and interactions. Depression showed the opposite difficulty. The best-performing method, extremely randomized trees, lowered RMSE from 3.19 to 3.11 percentage points, with a held-out R^2 of 0.427. Boosting ranked first for frequent mental distress, frequent physical distress, and high cholesterol; extremely randomized trees also ranked first for disability. The average ranking therefore conceals several outcome-specific exceptions.

Aim 3. Which predictors benefited from a curved term?

The largest average gains from restoring a curved term were for hispanic or latino residents (1.03%), median household income (1.02%), bachelor’s degree or higher (0.99%). These comparisons ask whether a single nonlinear term helps after other predictors have already been modeled flexibly.

Table 3

Gains compare the full additive model to the same model with one predictor constrained to a line. Negative values favor a linear term. All 150 predictor–outcome comparisons and paired state intervals appear in the workbook.

Predictor	Mean RMSE gain (%)	Improved outcomes / 15
Hispanic or Latino residents	1.03	12
Median household income	1.02	13
Bachelor’s degree or higher	0.99	15
Adults aged 45–64	0.47	15
Adults aged 65 or older	0.29	13
Non-Hispanic White residents	0.22	11
Civilian employed	0.14	11
Non-Hispanic Black residents	0.11	5
Residents at or above poverty threshold	0.04	10
Female residents	-0.07	5

The predictor ranking also depended on the outcome (Fig. 1). Allowing curvature in the percentage of Hispanic or Latino residents yielded its largest observed gain for high blood pressure (3.0%). Median household income showed its largest gain for COPD (2.3%); college attainment showed its largest gain for fair or poor health (3.0%). These are exploratory maxima across 150 comparisons. The terms describe the composition of places and should not be read as effects on individual residents.

Age composition and interactions

Across outcomes, adding the two quadratic age terms reduced error by 0.86% on average; using splines for both age percentages reduced it by 1.09%. For cancer, the reductions were 3.59% and 4.10%, respectively. Among all 45 individual products, college attainment by percentage of White residents ranked first for average incremental prediction. Percentage aged 65 or older by percentage of White residents ranked second, and percentage aged 65 or older by percentage of Black residents ranked fourth. Thus age appeared in several of the strongest individual interactions, although its predictive contribution depended on what it was paired with.

Table 4
Incremental predictive value of restoring interactions involving tract age composition

Terms restored	Mean reduction in RMSE	Outcome with largest reduction
All 17 age products	0.85%	Cancer, 2.99%
Age × racial and ethnic composition	0.29%	Cancer, 1.98%
Age × socioeconomic indicators	0.29%	Stroke, 0.74%
Age × female percentage	0.03%	Frequent mental distress, 0.22%
Percentage aged 45–64 × percentage aged 65+	0.07%	Cancer, 0.20%

Note. Comparisons retain all other terms. Age percentages describe tract populations. The 17 products comprise six age-by-composition, eight age-by-socioeconomic, two age-by-sex-composition, and one age-by-age product. Individual rankings and quadratic comparisons are in the age-analysis workbook.

Discussion

The central finding was that much of the benefit commonly attributed to machine learning could be obtained with a flexible regression. All four alternatives improved on the linear model for each outcome. Yet the model with the best mean performance used a small set of recognizable terms: a curve for each area characteristic and products for pairs of characteristics. A researcher could inspect these terms and plot predicted prevalence for realistic combinations of predictors. The tree comparisons remain valuable because they show where greater flexibility was useful beyond that specification.

The largest absolute R^2 and the largest relative improvement answered different questions. For cancer, a researcher choosing a prediction method should ask how much error remains after using basic area composition. For diabetes, the observed gain suggests that the functional form matters when approximating PLACES estimates in other states. Even a reproducible gain could arise from nonlinearities in the small-area estimation procedure, correlations among inputs, or geographic differences in their distributions.

The method that ranks first depends partly on what it is asked to predict. Within a familiar setting, a model can exploit features of data collection that correlate with the outcome. A study of pneumonia detection showed how performance changed across hospital systems [21]. That example concerns images, but the underlying validation question applies here: which relationships survive a change in setting? Leaving states out tests one form of geographic transfer. Because training and validation tracts still receive outcomes from the same PLACES estimation framework, the test cannot show whether the fitted relationships transfer to independent surveillance records. This is why agreement with published tract estimates and accuracy against local measurements warrant separate evaluation.

Additive splines reduced average RMSE by 6.0% versus 7.8% after adding products. Extra trees and boosting achieved average gains of 5.1% and 7.4%, respectively. These contrasts help distinguish smooth curvature from the additional predictive flexibility available to interaction models. They cannot quantify a biological contribution from interactions. In particular, curvature in the percentage of residents in a racial or ethnic category is a pattern in area composition and modeled prevalence, not a property of individuals.

For predictor interpretation, retaining versus removing curvature answers a more specific question than variable-importance rankings. A predictor may be essential to accurate prediction and still show little gain from a curve if its relationship is nearly linear conditional on other variables. Conversely, a modest gain from curving income can recur across several outcomes without making income their strongest predictor. The income, Latino population percentage, and education results are best treated as candidates for subsequent examination using fitted curves, adequate local sample sizes, and independently measured outcomes.

The age analysis also changes how the predictor rankings should be read. The modest average gain from giving just one age percentage a curved term understated what the two age percentages contributed jointly for cancer. An interaction with racial and ethnic composition can improve prediction for several reasons, including correlations between demographic composition and the inputs used to generate a prevalence estimate. Neither a high rank nor a reduction in error identifies biological effect modification. Interpretation also depends on the unit: two tracts with similar percentages aged 65 or older can differ in the distribution of ages within that group. Individual age in years was unavailable.

Strengths and limitations

Using identical geographic folds, a common linear comparator, and all 15 outcomes limits the opportunity for a model to appear superior because it was evaluated on easier tracts. Whole-state holdouts are a stronger geographic test than randomly withholding neighboring tracts. Nevertheless, adjoining states can share regional characteristics. The state-resampling intervals quantify variation among test states conditional on the fitted predictions. They do not establish that the narrow average difference between the two leading models would persist after refitting or retuning in another sample.

Choosing a target and choosing an algorithm are separate sources of risk. Obermeyer and colleagues showed that a health-care algorithm could reproduce racial inequity when health spending was used as a proxy for need [22]. In our application the target is the published prevalence estimate, and income and demographic predictors overlap with inputs to the target. This does not invalidate the comparison between models; it limits what the comparison establishes. Before using predictions to direct services, a health department would need to assess calibration within its own jurisdictions [23], check for systematic underprediction in underserved communities, and decide whether prevalence alone measures the need it intends to address.

PLACES uses multilevel regression and poststratification to generate tract estimates [3]. The inclusion of related demographic and socioeconomic variables in the source estimates means that prediction error cannot be interpreted as error against directly observed tract prevalence. Historical tract matching also excluded areas with changed identifiers, and complete-case selection can limit geographic coverage. We evaluated fixed model configurations in one set of state folds. The per-outcome winner and largest predictor gain were selected using those same results. Prospective validation should preselect the candidate model, assess calibration by geography and population composition, and compare predictions with independently recorded outcomes.

Prediction studies can also perform well on average while remaining difficult to evaluate or reproduce. Guidance on reporting prediction models emphasizes clear definitions of the population, predictors, outcome, and validation procedure [24]. Risk-of-bias assessment adds attention to selection and analytical choices [25]. Here, tract exclusions, a single set of geographic folds, and post hoc selection of the lowest-error model all affect interpretation. Our state intervals quantify variation in the observed validation errors; they do not reflect all the uncertainty involved in selecting an algorithm. Reporting the complete outcome grid helps readers assess whether an apparent advantage is broad or concentrated in a few outcomes.

Implications. A health department choosing among these methods could compare the expected improvement for its target outcome against the convenience of a transparent additive model. A next validation would compare predictions with geographically independent survey measurements or administrative health records in harmonized tract boundaries. Calibration in the intended deployment area, and error patterns across socioeconomic and demographic groups, merit evaluation before a model is used to allocate resources.

For these aggregate data, the practical choice is between a transparent regression that performed well across many outcomes and a more flexible ensemble that occasionally improved prediction for a particular condition. Work on interpretable modeling argues for evaluating understandable alternatives when predictions inform consequential decisions [26]. Clinical ML implementation studies also emphasize workflow, monitoring, and local validation [27]. The corresponding public health questions are concrete: how large is the improvement in prevalence points, which communities experience larger errors, and would that improvement change a service decision?

Conclusion

For prediction in held-out states, regression with curves and pairwise interactions was the strongest overall benchmark; boosting was the strongest tree-based method. Diabetes, stroke, and COPD benefited most from added flexibility, whereas cancer was the most predictable outcome overall. The value of allowing income, education, or demographic composition to curve varied by outcome. Public health researchers choosing a method should consider both the reduction in error and its size in prevalence percentage points for the condition they intend to study.

Declarations

Competing interests.

The author declares no competing interests.

Ethics and consent.

This secondary analysis used publicly available, aggregate census tract estimates. There was no participant contact or access to identifiable individual records. Individual informed consent was inapplicable.

Use of artificial intelligence.

OpenAI's ChatGPT assisted with analysis (statistical programming, model estimation, and preparation of tables and figures) and manuscript editing only; it was not used for manuscript writing. Assari supervised this process.

Funding.

Shervin Assari receives partial support through National Institutes of Health (NIH) grants U54MD007598, S21MD018960, and R25MD007610 to Charles R. Drew University of Medicine and Science. The content is the responsibility of the author and does not necessarily represent the official views of the NIH.

Author contributions.

Shervin Assari conceived the study, directed the analyses, interpreted the findings, and revised the manuscript.

Data availability.

The analysis used public 2014–2018 American Community Survey five-year estimates and the 2025 CDC PLACES census tract release. Supplementary Appendix A contains model performance, state folds, and predictor curvature results. Supplementary Appendix B contains the age analyses and all 45 interaction rankings. Both workbooks, the aggregate analytic dataset, analysis scripts, and replication instructions are provided in the accompanying supplementary ZIP.

References

1. Maharana A, Nsoesie EO (2018) Use of Deep Learning to Examine the Association of the Built Environment With Prevalence of Neighborhood Adult Obesity. JAMA Netw Open. ;1(4):e181535. 10.1001/jamanetworkopen.2018.1535. PMID: 30646134
2. Jean N, Burke M, Xie M, Alampay Davis WM, Lobell DB, Ermon S (2016) Combining satellite imagery and machine learning to predict poverty. Science. ;353(6301):790–794. 10.1126/science.aaf7894. PMID: 27540167
3. Centers for Disease Control and Prevention PLACES methodology. https://www.cdc.gov/places/methodology/index.html
4. Centers for Disease Control and Prevention About PLACES: Local Data for Better Health. https://www.cdc.gov/places/about/index.html
5. Altman DG, Royston P (2006) The cost of dichotomising continuous variables. BMJ. ;332(7549):1080. 10.1136/bmj.332.7549.1080. PMID: 16675816
6. Greenland S (1995) Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology. ;6(4):356–365. 10.1097/00001648-199507000-00005. PMID: 7548341
7. Knol MJ, VanderWeele TJ (2012) Recommendations for presenting analyses of effect modification and interaction. Int J Epidemiol. ;41(2):514–520. 10.1093/ije/dyr218. PMID: 22253321
8. VanderWeele TJ, Tchetgen Tchetgen EJ (2014) Attributing effects to interactions. Epidemiology. ;25(5):711–722. 10.1097/ede.0000000000000096. PMID: 25051310
9. Wiemken TL, Kelley RR (2020) Machine Learning in Epidemiology and Health Outcomes Research. Annu Rev Public Health. ;41:21–36. 10.1146/annurev-publhealth-040119-094437. PMID: 31577910
10. Obermeyer Z, Emanuel EJ (2016) Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med. ;375(13):1216–1219. 10.1056/nejmp1606181. PMID: 27682033
11. Rajkomar A, Dean J, Kohane I (2019) Machine Learning in Medicine. N Engl J Med. ;380(14):1347–1358. 10.1056/nejmra1814259. PMID: 30943338
12. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med. ;25(1):44–56. 10.1038/s41591-018-0300-7. PMID: 30617339
13. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR (2016) Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. ;316(22):2402–2410. 10.1001/jama.2016.17216. PMID: 27898976
14. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature. ;542(7639):115–118. 10.1038/nature21056. PMID: 28117445
15. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol. ;110:12–22. 10.1016/j.jclinepi.2019.02.004. PMID: 30763612

16. US Census Bureau 2014–2018 American Community Survey five-year detailed tables. https://www.census.gov/programs-surveys/acs/data/summary-file.2018.html
17. Roberts DR, Bahn V, Ciuti S et al (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913–929. 10.1111/ecog.02881
18. Hastie T, Tibshirani R (1986) Generalized additive models. Stat Sci 1(3):297–310. 10.1214/ss/1177013604
19. Geurts P, Ernst D, Wehenkel L (2006) Extremely randomized trees. Mach Learn 63:3–42. 10.1007/s10994-006-6226-1
20. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232. 10.1214/aos/1013203451
21. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK (2018) Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. ;15(11):e1002683. 10.1371/journal.pmed.1002683. PMID: 30399157
22. 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. 10.1126/science.aax2342. PMID: 31649194
23. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW (2019) Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. ;17(1):230. 10.1186/s12916-019-1466-7. PMID: 31842878
24. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P (2024) TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. ;385:e078378. 10.1136/bmj-2023-078378. PMID: 38626948
25. Wolff RF, Moons KGM, Riley RD et al (2019) PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med. ;170(1):51–58. 10.7326/M18-1376. PMID: 30596875
26. Rudin C (2019) Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nat Mach Intell. ;1(5):206–215. 10.1038/s42256-019-0048-x. PMID: 35603010
27. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D (2019) Key challenges for delivering clinical impact with artificial intelligence. BMC Med. ;17(1):195. 10.1186/s12916-019-1426-2. PMID: 31665002

Figures

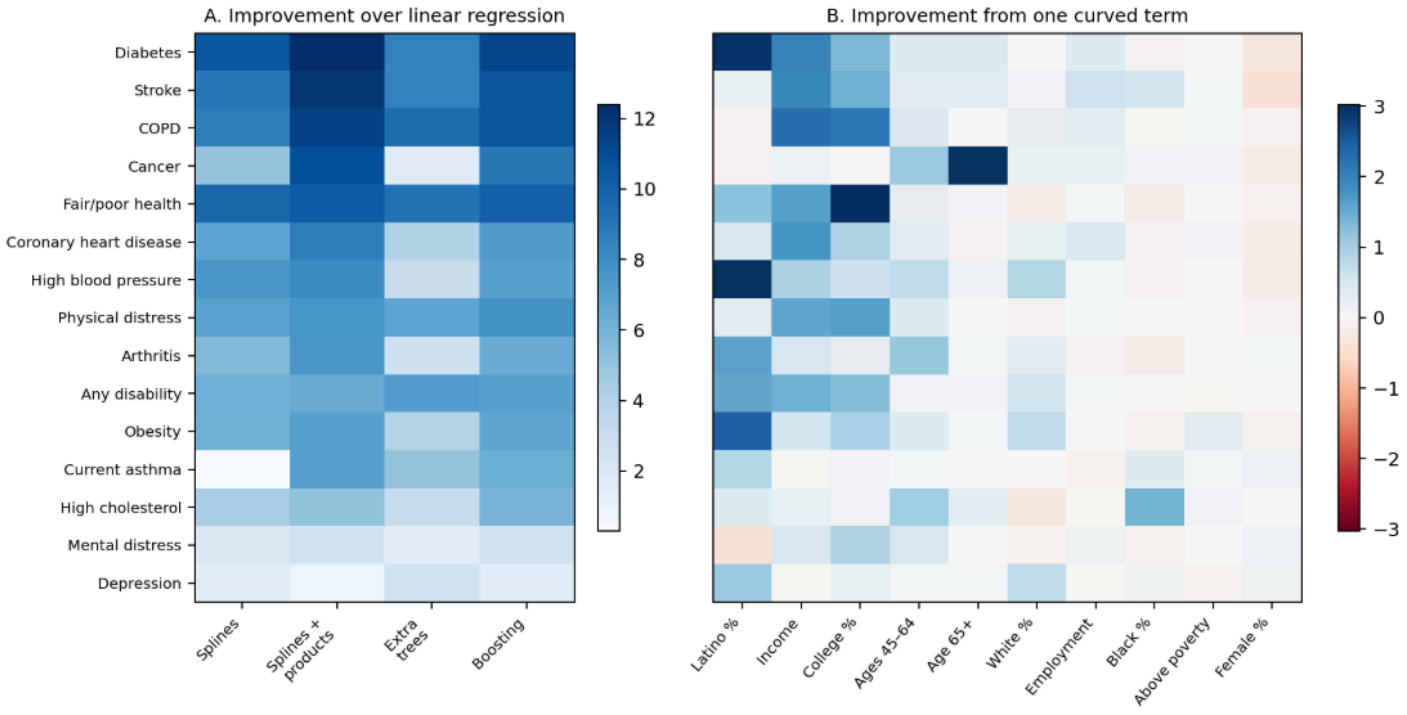

Figure 1

Relative reduction in held-out RMSE (%)

Panel A compares each model with linear regression. Panel B shows the gain from retaining one predictor's curved term with the other nine already modeled flexibly. Darker blue denotes larger gains; red in panel B denotes worse prediction. Scales differ between panels.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- NeighborhoodHealthPredictionSupplement.zip