


RESEARCH ARTICLE

Open Access



Incorporating machine learning and social determinants of health indicators into prospective risk adjustment for health plan payments

Jeremy A. Irvin^{1*} , Andrew A. Kondrich¹, Michael Ko², Pranav Rajpurkar¹, Behzad Haghgoo¹, Bruce E. Landon^{3,4}, Robert L. Phillips⁵, Stephen Petterson⁶, Andrew Y. Ng¹ and Sanjay Basu^{4,7,8}

Abstract

Background: Risk adjustment models are employed to prevent adverse selection, anticipate budgetary reserve needs, and offer care management services to high-risk individuals. We aimed to address two unknowns about risk adjustment: whether machine learning (ML) and inclusion of social determinants of health (SDH) indicators improve prospective risk adjustment for health plan payments.

Methods: We employed a 2-by-2 factorial design comparing: (i) linear regression versus ML (gradient boosting) and (ii) demographics and diagnostic codes alone, versus additional ZIP code-level SDH indicators. Healthcare claims from privately-insured US adults (2016–2017), and Census data were used for analysis. Data from 1.02 million adults were used for derivation, and data from 0.26 million to assess performance. Model performance was measured using coefficient of determination (R^2), discrimination (C-statistic), and mean absolute error (MAE) for the overall population, and predictive ratio and net compensation for vulnerable subgroups. We provide 95% confidence intervals (CI) around each performance measure.

Results: Linear regression without SDH indicators achieved moderate determination (R^2 0.327, 95% CI: 0.300, 0.353), error (\$6992; 95% CI: \$6889, \$7094), and discrimination (C-statistic 0.703; 95% CI: 0.701, 0.705). ML without SDH indicators improved all metrics (R^2 0.388; 95% CI: 0.357, 0.420; error \$6637; 95% CI: \$6539, \$6735; C-statistic 0.717; 95% CI: 0.715, 0.718), reducing misestimation of cost by \$3.5 M per 10,000 members. Among people living in areas with high poverty, high wealth inequality, or high prevalence of uninsured, SDH indicators reduced underestimation of cost, improving the predictive ratio by 3% (~\$200/person/year).

Conclusions: ML improved risk adjustment models and the incorporation of SDH indicators reduced underpayment in several vulnerable populations.

Keywords: Risk estimation, Machine learning, Social determinants of health

* Correspondence: jirvin16@cs.stanford.edu

¹Department of Computer Science, Stanford University, 353 Serra Mall, Stanford, CA 94305, USA

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Background

Public and private regulators use risk adjustment models to prevent adverse selection, anticipate budgetary reserve needs, and offer care management services to high-risk individuals [1]. Preventing risk selection by insurers is a critical ethical, legal, and societal goal that risk adjustment models can address. Risk adjustment models attempt to capture the relationship between demographic and clinical variables (risk adjusters) and subsequent healthcare utilization or spending. The models are commonly derived through standard linear regression methods or their extensions, and rely on individual-level data commonly captured in administrative claims datasets [2]. All of the available models on the current commercial market are linear or log-linear regression models that leverage the same basic elements such as age, sex, diagnostic and procedure codes [3].

Risk adjustment modeling may be improved by both methodological and conceptual advances in the risk modeling and healthcare services literature. From a methodological standpoint, newer machine learning methods have recently emerged as alternatives or complements to linear regression for predicting highly variable health outcomes using large sparse datasets, including estimating healthcare costs using claims data [4, 5]. While traditional risk adjustment models are limited in modeling complexity and tend to underpredict expenditures of populations with very high expenditures [6, 7], machine learning methods may help to capture complex non-linear relationships and interaction terms among variables, which could explain why some individuals with complex constellations of risk factors and diagnoses experience substantially higher cost than predicted. For example, among people with low income and diabetes receiving insulin, food insecurity is associated with hypoglycemia and emergency room visits during the last week of each month (after income from a first-of-the-month paycheck is deprived) and hypoglycemic medications are still being taken [8]. These complex relationships are hard to model in standard risk equations, but can be potentially better captured by interactions-focused, nonlinear machine learning algorithms. Despite the promise of machine learning for risk adjustment, machine learning techniques have not yet been widely adopted for risk adjustment. This is partially because the machine learning models developed to date have not yet demonstrated superior predictive performance over traditional linear models on large datasets with more than a million enrollees [2].

From a conceptual standpoint, risk adjustment may also be improved by including additional area-level indicators of social determinants of health (SDH), such as poverty, unemployment, and education, which contribute to risk, utilization and cost [9–11]. Since before the

UK Black Report and the Health Divide, epidemiologists have shown that while cultural and individual behavioral choices influence health, living conditions including the availability of resources (e.g., clean air and water), working conditions, and quality of food and housing have a particularly profound association with health outcomes [12]. More recent initiatives to directly address these ‘social determinants’ of health include strategies to refer patients with food insecurity to food pantries, those that are homeless to direct housing resources, and those with challenges with transportation to assisted transport services, as a means to improve clinical outcomes such as nutrition-related chronic disease metrics (e.g., nutrition affecting blood pressure and diabetes glycemic control) and to improve the ability to access healthcare visits and reduce stress-related adverse health outcomes [13].

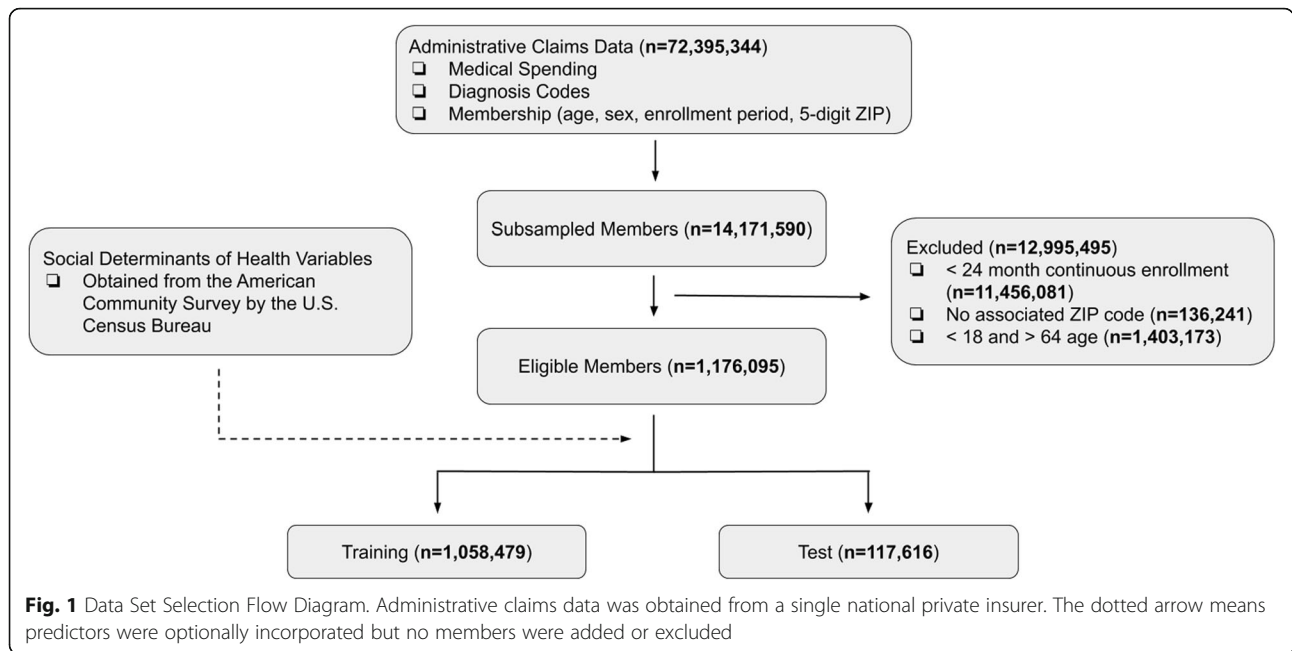
The inclusion of SDH indicators into risk adjustment may particularly help plan payment estimation. SDH indicators may help capture previously unmeasured factors that could influence the course of disease, such as how poverty may affect chronic disease outcomes by affecting the ability to pay for medications or more nutritious foods, or how unemployment relates to mental health and associated course of disease related to depression and lower adherence [14, 15]. Individual-level SDH factors are rarely assessed or included in commonly-available data, but area-level SDH indicators are readily assessed by national data sources [16], and may be linked to the 5-digit ZIP code often available in claims data. Area-level SDH indicators were recently incorporated into risk adjustment models for the Massachusetts Medicaid program; their inclusion improved concurrent annual healthcare spending predictions for low-income adults [17]. It remains unclear, however, to what extent incorporating area-level SDH indicators could improve prospective annual healthcare spending predictions, particularly for the privately-insured population who constitute the largest share of insured people in the US, but for whom SDH factors may be less visible or influential than for the Medicaid population.

The objective of this study was to assess whether prospective risk adjustment models may be improved by machine learning methods and by the incorporation of area-level SDH indicators in a national privately-insured adult population.

Method

Data

Our primary data were healthcare claims from a single large national commercial insurer operating in all 50 US states, Washington D.C., and Puerto Rico (Fig. 1). From the claims data, we included privately-insured individuals 18 through 64 years old who had at least 24 months of continuous enrollment. Individuals who switched



plans were implicitly excluded due to the continuous enrollment criteria, but individuals who moved were included. We used demographics (age, sex) and diagnostic codes (Clinical Classification Software categories [18]) as candidate risk adjustment variables at the individual level, and SDH indicators at the 5-digit ZIP code-level from the American Community Survey (ACS) by the U.S. Census Bureau (Table 1) [16]. SDH indicators were

selected to reflect current conceptual theories concerning a broad range of social, economic, and health system factors that may influence health risk, utilization, or cost (Supplementary Information Table S1) [19, 20]. The resulting dataset contained claims data from 1.18 million unique members, which we randomly partitioned into training data (1.06 million members) for model derivation and test data (0.12 million members) for model

Table 1 Statistics of ZIP Code-Level Social Determinants of Health

SDH Variable	Mean [Median] (Std)
Median Income in the Past 12 Months, \$	26,546 [25727] (6230)
Families Under 0.5 Ratio of Income to Poverty Level in the Past 12 Months, %	4.7 [4.3] (2.7)
Families Between 0.5 and 0.74 Ratio of Income to Poverty Level in the Past 12 Months, %	3.2 [3.0] (1.7)
Families Between 0.75 and 0.99 Ratio of Income to Poverty Level in the Past 12 Months, %	3.6 [3.4] (1.6)
Families Received Food Stamps/Snap in the Past 12 months, %	14.2 [13.6] (6.6)
Population Unemployed, %	5.4 [5.2] (1.9)
Gini Index of Income Inequality	45.2 [45.1] (3.6)
Population Obtained High School Diploma, %	43.0 [42.9] (4.8)
Population Obtained Bachelor's Degree, %	16.1 [15.2] (6.2)
Population Speak English Less than "Very Well", %	10.5 [5.6] (12.5)
Families with Single Parent, %	22.9 [22.7] (6.2)
Population Without Health Insurance Coverage, %	11.3 [10.6] (4.9)
Population African American, %	9.9 [4.5] (13.5)
Population Asian, %	2.9 [1.2] (4.7)
Population American Indian and Alaska Native, %	1.5 [0.3] (6.4)
Population Hispanic or Latino, %	11.9 [5.5] (15.8)
Population White, %	71.8 [77.7] (22.3)

SDH variables were obtained from the 2012–2016 American Community Survey 5-year estimates from the U.S. Census Bureau

performance assessment [21]. There was no overlap of individuals among the two data subsets. Demographic statistics of the data subsets by geographic location are reported in Supplementary Information Table S2.

Outcome

We sought to prospectively predict 2017 individual-level total annual healthcare spending from 2016 data. As a secondary objective, we also considered concurrent risk adjustment, predicting 2016 member-level annual spending from 2016 data, as is done, for instance, in the Affordable Care Act health insurance exchanges (see [Supplementary Information](#)). We estimated total annual spending by summing standardized costs in U.S. Dollars over 12 months, including post-year claims corrections, and including zero spending among enrolled individuals without medical claims. To inhibit outliers from affecting model fit, costs in the training set were top-coded at \$400,000 (cost larger than \$400,000 was replaced with \$400,000), which corresponded to the top 0.1% cost of members in the training set. Top-coding is performed to reduce model sensitivity to skewness and kurtosis, and has been preferred over dropping members with high cost since these cases can be indicative of specific conditions which are associated with high cost [2, 22, 23].

Model development

A 2-by-2 factorial design was employed to compare modeling approaches (linear regression versus the machine learning approach of gradient boosted decision trees), and variable choice (demographics and diagnostic codes alone versus additional area-level SDH indicators). In each of the methods, individual-level predictors with their associated area-level predictors are input to the model together as if they were all individual-level properties.

Linear regression approach

A linear model derived through ordinary least squares regression was trained to predict 2017 spending based on 2016 member characteristics. We additionally developed penalized linear regression models using methods that may better address collinearity (Least Absolute Shrinkage and Selection Operator [LASSO]), as detailed in the [Supplementary Information](#). LASSO regression tends to sparsely select among collinear variables by forcing coefficients to zero for all but one of the collinear variables [24, 25].

Machine learning approach

The machine learning approach investigated in this study was gradient boosted decision trees [26]. This approach involves the construction of an ensemble of decision trees, where each tree learns from the errors of the

prior tree (a “boosting” approach) to iteratively improve predictions [27]. With each iteration, a new tree is constructed by sampling from the data and identifying which variable most effectively divides the members into groups with low within-group variation in cost and high between-group variation in cost. This variable selection process is repeated to further divide each resulting subset of the data, producing a series of branches in the decision tree. The tree is added to the current ensemble, and then the next tree is fit using the same process on the residuals of the ensemble.

We chose gradient boosted decision trees over alternative machine learning methods because the approach has been shown to handle mixes of categorical and continuous covariates, capture nonlinear relationships, and scale well to large amounts of data [28]. Moreover, it is straightforward to obtain variable importance rankings from the model, which may permit the approach to be more interpretable than many other machine learning methods, for which acceptability in a healthcare services context may critically depend on visualizing “black box” predictions [29]. We used the LightGBM framework to develop the models, which implements several algorithmic optimizations on standard gradient boosting to allow for additional training efficiency [30]. A detailed treatment of gradient boosted decision trees and LightGBM is provided in the [Supplementary Information](#). We used 3-fold cross validation on the training data subset to select the parameters for the model, including the number of trees, the maximum depth of each tree, and the minimum level of loss reduction necessary to partition leaf nodes, based on which achieved the lowest mean squared error averaged across the 3 folds [21]. We then refitted the model to the full training set using the best parameters determined from 3-fold cross validation, which can further help reduce overfitting. We additionally developed random forest and shallow multilayer perceptron models using a similar training procedure, as detailed in the [Supplementary Information](#) [31, 32].

Model testing and statistical analysis

We evaluated the performance of the prospective risk adjustment models on the test set. The performance metrics are detailed below.

Goodness of fit

We evaluated model goodness of fit using the coefficient of determination (R^2) and the mean absolute error (MAE). We estimated the R^2 with confidence intervals using the nonparametric bootstrap with 5000 bootstrap replicates [33], and the MAE with confidence intervals using a paired t-test.

Discrimination

We assessed discrimination using the concordance-statistic (*C*-statistic), a rank correlation metric for assessing the model's ability to order members by their spending [34]. The *C*-statistic estimates the probability that, for a randomly selected pair of members, the member with the higher cost will be correctly predicted as having higher cost by the model [35]. The *C*-statistic is the generalization of the area under the receiver operating characteristic curve from the binary to the continuous outcome setting, where a result between 0.7 and 0.8 is considered acceptable, between 0.8 and 0.9 is considered good, and above 0.9 is considered excellent [36]. We estimated confidence intervals for the *C*-statistic using a jack-knife procedure [37].

Subgroup analyses

Risk adjustment models often underpredict spending for specific subgroups of enrollees leading to underpayment to the insurer, and there is evidence that insurers explicitly make health plans less desirable for enrollees in undercompensated groups [38, 39]. To evaluate the performance of the models on vulnerable subgroups, we defined test data subgroups using age, sex, and area-level SDH indicators. SDH indicator subgroups included individuals living in ZIP codes in the lowest decile of household income; lowest decile of education level (by high school diploma and by bachelor's degree receipt); highest decile of Gini index for inequality; low ratio of income to poverty level; high proportion of households receiving food stamps; high proportion with single parents; high unemployment; high uninsurance rate; and high proportion reporting they do not speak English "very well" (see Supplementary Information Table S1 for decile thresholds).

The performance of the model on each subgroup was measured using the predictive ratio [40] and net compensation [39, 41]. The predictive ratio for a subgroup was computed as the ratio of the mean of observed spending to the mean of predicted spending over the subgroup, where a value above 1 indicates underestimation of cost and a value below 1 indicates overestimation [17]. We estimated 95% confidence intervals around the predictive ratio using the delta method [42]. Net compensation was used as a measure on the dollar scale, and was computed as the mean difference between predicted spending and observed spending over the subgroup, where a value below 0 indicates underestimation of cost and a value over 0 indicates overestimation. We estimated 95% confidence intervals around the net compensation values using a paired *t*-test.

Analyses were approved by the Stanford Institutional Review Board (eProtocol #42334), and performed in Python version 3.6.6 [43] and R version 3.5.0 [44], using

the code shared online for reproducibility at: <https://github.com/stanfordmlgroup/risk-adjustment-ml>.

Results

Descriptive statistics on the data subsets are detailed in Table 2. The test set had a mean age of 41.1 years (median 41.0; IQR 30.0, 53.0) and was 48.9% female. Top-coding cost at \$400,000 eliminated approximately 2.8% of dollars and test set members had a mean top-coded annual healthcare cost of \$6677 (median 855; IQR 161, 3847). Around 17.7% of members in the test set had zero annual healthcare cost.

Table 3 shows the test set performance of the prospective linear and machine learning models without and with the SDH indicators.

Linear regression without SDH indicators

The linear regression model without SDH indicators, when derived through ordinary least squares regression, had the largest standardized coefficients (indicating highest importance among covariates in the model) for age and sex indicators and diagnostic coding for birth complications and chronic kidney disease (see Supplementary Information Table S3). The model had a R^2 of 0.327 (95% CI 0.300, 0.353), MAE of \$6992 (95% CI 6889, 7094), and *C*-statistic of 0.703 (95% CI 0.701, 0.705). Linear models derived through LASSO had similar performance metrics but tended to favor diagnoses more than traditional least squares (see Supplementary Information Tables S3 and S5).

Linear regression with SDH indicators

The inclusion of SDH indicators in the linear regression model had no substantial effect on the overall performance metrics. The model had a R^2 0.327 (95% CI 0.300, 0.354), MAE of \$6991 (95% CI 6889, 7094), and *C*-statistic of 0.700 (95% CI 0.699, 0.702).

Machine learning without SDH indicators

Switching from a linear regression model to the machine learning model significantly improved determination, significantly reduced error, and significantly improved discrimination. Specifically, the machine learning model without SDH indicators had a R^2 of 0.388 (95% CI 0.357, 0.420), MAE of \$6637 (95% CI 6539, 6735), and *C*-statistic of 0.717 (95% CI 0.715, 0.718). The multilayer perceptron and random forest models outperformed the linear models but performed worse than the LightGBM model across all metrics (Supplementary Information Table S5).

Machine learning with SDH indicators

The inclusion of SDH indicators in the machine learning model also had no substantial effect on the overall

Table 2 Characteristics of Members in the Dataset Subsets

Characteristic	Training Set	Test Set
Members Total, No.	1,058,479	117,616
Female Total, No. (%)	517,364 (48.9%)	57,469 (48.9%)
Members from ZIP codes without measured SDH variables ^a , No. (%)	1074 (0.1%)	115 (0.1%)
Population statistics, mean [median] (SD)		
Age, y	41.1 [41.0] (13.1)	41.1 [41.0] (13.1)
2017 Annual Cost, \$	6946 [861] (28,240)	6868 [855] (27,826)
2017 Top-coded Annual Cost ^b , \$	6762 [861] (23,822)	6677 [855] (23,536)

The training set was used to develop the models and the test set was used to evaluate the models

^aThe SDH variables of these members were imputed with the median values of SDH variables over all ZIP codes, and an additional indicator variable was used to identify whether members fall into this category

^bStatistics of cost when top-coding at \$400,000 (values higher than \$400,000 were replaced with \$400,000)

performance metrics above the machine learning model without SDH indicators. The model had a R² of 0.387 (95% CI 0.357, 0.419), MAE of \$6634 (95% CI 6536, 6732), and C-statistic of 0.716 (95% CI 0.714, 0.717). We created variable importance rankings to assist in the interpretation of the machine learning model. Diagnosis predictors had the largest importance metrics in the machine learning model, with the most important predictors being chronic kidney disease, deficiency and other anemia, and other aftercare (see Supplementary Information Table S4).

Subgroup analyses

Table 4 compares the predictive ratios and net compensation values for the machine learning model without and with SDH indicators. The addition of SDH indicators resolved or reduced underestimation of risk on all of the SDH-based subgroups, but the 95% confidence intervals

were overlapping between the non-SDH and SDH-including models among all subgroups. On one of the high-poverty subgroups, the subgroup with a high proportion of non-fluent English speakers, the subgroup with a high prevalence of uninsured, and the subgroup of individuals who lived in areas with a large proportion of households on food stamps, the incorporation of SDH indicators resolved the underestimation of risk. Among subgroups of individuals who lived in areas with high poverty, high wealth inequality, and high prevalence of uninsured, the machine learning model trained with SDH indicators substantially reduced underestimation of cost among the subgroup, improving the predictive ratio by 3% (and net compensation by \$200 per person) over the model trained without SDH indicators. The addition of SDH indicators led to small additional overpayment on the 4 subgroups for which the model without SDH indicators did not substantially underestimate risk (predictive ratio < 1.01), specifically one of the high-poverty subgroups, the subgroup with a large unemployed population, the subgroup with a low percentage of high school graduates, and the subgroup with a large number of single-parent families. Additional subgroup analyses among all models are presented in Supplementary Information Tables S6, 7, 8.

Table 3 Performance Measures of the Prospective Linear and Machine Learning Models on the Test Set

Evaluation Metric	No SDH	SDH
R ² (95% CI) ^a		
Linear	0.327 (0.300, 0.353)	0.327 (0.300, 0.354)
ML	0.388 (0.357, 0.420)	0.387 (0.357, 0.419)
MAE (95% CI) ^b		
Linear	6992 (6889, 7094)	6991 (6889, 7094)
ML	6637 (6539, 6735)	6634 (6536, 6732)
C-statistic (95% CI) ^c		
Linear	0.703 (0.701, 0.705)	0.700 (0.699, 0.702)
ML	0.717 (0.715, 0.718)	0.716 (0.714, 0.717)

Comparison of performance measures between linear regression and machine learning prospective risk adjustment models, predicting 2017 yearly top-coded spending from 2016 characteristics. The SDH model additionally includes SDH variables obtained from U.S. Census data (see Table 1)

^aConfidence intervals for R² were constructed using the nonparametric bootstrap [21]

^bConfidence intervals for MAE were constructed using a paired t-test

^cConfidence intervals for C-statistic were constructed using a jackknife procedure [25]

Additional results

Binned scatter plots of the prospective risk adjustment models on the test set are shown in Fig. S1. We additionally explored the effect of using binary diagnosis predictors instead of counts (Supplementary Information Table S9), the effect of top-coding cost (Supplementary Information Table S10), the effect of including lab results (Supplementary Information Table S11), and the development of concurrent risk adjustment models (Supplementary Information Table S12).

Discussion

We observed that switching from a linear regression model to a gradient boosting ML model significantly improved determination and discrimination and reduced

Table 4 Predictive Ratio and Net Compensation Values of Prospective Machine Learning Models on SDH-Based Subgroups in the Test Set

Subgroup	No. (%)	2017 Spending (\$) ^a	Model Predictive Ratio ^b and Net Compensation ^c	
			ML (95% CI)	ML with SDH (95% CI)
Total	117,616 (100)	6677	1.000 (0.976, 1.024) 0 (− 105, 105)	1.000 (0.976, 1.024) 0 (− 105, 105)
Poverty				
Median Income in the Past 12 Months, \$	4923 (4.2)	10,818	1.017 (0.915, 1.120) − 183 (− 836, 470)	1.006 (0.905, 1.108) − 67 (− 729, 595)
Families Under 0.5 Ratio of Income to Poverty Level in the Past 12 Months, %	7932 (6.7)	9344	0.966 (0.882, 1.050) 331 (− 138, 801)	0.948 (0.865, 1.031) 510 (33, 987)
Families Between 0.5 and 0.74 Ratio of Income to Poverty Level in the Past 12 Months, %	6651 (5.7)	8952	1.010 (0.912, 1.108) − 89 (− 599, 420)	0.988 (0.892, 1.084) 109 (− 408, 627)
Families Between 0.75 and 0.99 Ratio of Income to Poverty Level in the Past 12 Months, %	7194 (6.1)	9395	1.052 (0.956, 1.148) − 467 (− 977, 43)	1.010 (0.919, 1.101) − 94 (− 613, 425)
Families Received Food Stamps/Snap in the Past 12 months, %	9009 (7.7)	9001	1.028 (0.941, 1.115) − 247 (− 684, 191)	0.996 (0.912, 1.079) 39 (− 409, 487)
Population Unemployed, %	10,278 (8.7)	7055	0.961 (0.886, 1.036) 289 (− 71, 649)	0.957 (0.882, 1.032) 316 (− 51, 683)
Gini Index of Income Inequality	16,155 (13.7)	6138	1.054 (0.985, 1.122) − 312 (− 578, − 46)	1.021 (0.955, 1.087) − 126 (− 393, 140)
Education				
Population Obtained High School Diploma, %	9482 (8.1)	7555	0.987 (0.900, 1.073) 102 (− 324, 529)	0.974 (0.889, 1.058) 205 (− 227, 637)
Population Obtained Bachelor’s Degree, %	4169 (3.5)	11,338	1.032 (0.923, 1.142) − 353 (− 1139, 433)	1.027 (0.917, 1.136) − 294 (− 1080, 492)
Other				
Population Speak English Less than “Very Well”, %	23,659 (20.1)	5453	1.023 (0.963, 1.083) − 124 (− 346, 98)	0.989 (0.932, 1.046) 61 (− 161, 283)
Families with Single Parent, %	9097 (7.7)	9880	0.993 (0.910, 1.076) 65 (− 397, 527)	0.978 (0.896, 1.060) 224 (− 246, 693)
Population Without Health Insurance Coverage, %	13,656 (11.6)	8333	1.066 (0.990, 1.142) − 516 (− 885, − 147)	0.990 (0.921, 1.059) 83 (− 287, 454)

Comparison of machine learning prospective risk adjustment models without and with the addition of SDH indicators as predictors (see Table 1 for a complete list of SDH indicators). The predictions for each model were adjusted so that the mean of the predictions over the total test population was equal to the mean of the actual costs, resulting in a predictive ratio of exactly 1.0 over the total test set population. Subgroups were composed of members in the lowest decile of ZIP codes with respect to the corresponding SDH variable (see Supplementary Information Table S1). Only socioeconomic variables are considered in this subgroup analysis, and results on age and sex subgroups are shown in the [Supplementary Information](#)

^aSpending included all healthcare utilization in 2017 of members with full enrollment in 2016 and 2017. Values larger than \$400,000 were replaced with \$400,000

^bPredictive ratio for a subgroup was computed as the ratio of the mean of observed to the mean of predicted spending over the subgroup. Approximate confidence intervals for predictive ratios were computed with the delta method [40]

^cNet compensation for a subgroup was computed as the mean difference between predicted and observed spending in the subgroup. Confidence intervals were estimated using a paired t-test

absolute error in cost. We also observed that the inclusion of SDH indicators at the ZIP code-level reduced underestimation of cost among people living in vulnerable areas.

Prior studies have separately investigated whether machine learning and the incorporation of SDH indicators can improve risk adjustment. The use of machine learning for prospective risk prediction in a previous study did not demonstrate substantial improvements over

linear regression for a privately-insured population [4]. However, the addition of SDH indicators has been shown to improve concurrent risk adjustment models, including Medicare Advantage Plan quality rankings, Medicare’s Hospital Readmissions Reduction Program penalties, and concurrent annual healthcare spending among a state Medicaid population [17, 45, 46]. In our study, the incorporation of SDH indicators reduced cost

underestimation in several vulnerable subgroups, even among a commercially-insured population. Improving predictions of cost within these subgroups is important in order to address persistent inequalities that lead to bias in the estimation of payment [47–49].

Our study has important limitations. First, the risk models developed here are unlikely to generalize well to populations outside the U.S. as well as to Medicaid or Medicare populations for whom risk adjustment models may be particularly consequential to avoid adverse selection and maintain competitive and fair markets. However, the methods employed in this study could be used in developing specific models for those populations. Second, similar to other machine learning methods, the modeling approach used in this study is more complex than traditional linear regression. Although this may confer an advantage due to the potential of preventing ‘cheating’, in that machine learning models may be less susceptible to up-coding behaviors intended to inflate risk estimates [2], the complexity might also contribute to difficulty to understand how and why the model made a certain decision [29]. Third, since risk adjustment models are developed on historical data, they tend to perpetuate inequality of past spending trends if no explicit adjustments are made to account for the endogeneity of spending. Prior work has investigated methods to develop fairer healthcare payment models through data manipulation and modeling changes [39, 41, 50], which can be pursued in future studies. Fourth, the SDH indicators used in this study are at the area-level which may lead to bias or ecological fallacy in the risk adjustment models. However, combining the claims data used in this work with individual-level socioeconomic status variables was prohibited for privacy reasons. Fifth, 5-digit ZIP codes are not as homogeneous as Census Tracts or Census Block Groups, which have been used in previous linear regression models assessing SDH-associated effects for Medicaid and Medicare populations [51]. The risk for this study is a potential underestimation of the contribution of SDH to risk models. However, ZIP code is more readily available in commercial claims datasets. Sixth, there remains debate about whether adding in SDH indicators may allow for poorer healthcare to persist in healthcare organizations serving predominantly lower-income populations, by compensating them more in value-based payment models that adjust not only for outcomes but also for lower income for instance, although recent studies suggest this will not necessarily mask hospital quality [52]. Seventh, one key challenge is to predict per-member utilization rather than cost. However, given that cost is a key concern for payers and often disproportionate to utilization due to negotiated contracts and geographic variations in cost, we modeled overall costs to help understand how much geographic

parameters such as social determinants and machine learning could capture the complexities related to payment.

In the future, our ML approach may be improved upon in several ways. It may be possible to take advantage of the temporality of the data, for example by including more than one year of medical history. Additionally, it may be possible to train a hybrid (concurrent and prospective) model to leverage the continuous nature of medical enrollment, utilization, and claims [53]. Finally, using highly parameterized models such as deep neural networks could better capture nonlinear interactions between covariates and scale to large claims datasets, at the expense of interpretability [54]. We have shared our code in an open source manner to enable others to reproduce and extend our methods to other datasets.

Conclusion

The results of the current study suggest that machine learning methods and the inclusion of area-level SDH indicators may improve prospective risk adjustment models in a commercially insured population. The SDH indicators were particularly useful for populations living in vulnerable areas, while the machine learning approach had a greater impact on overall performance, leading to improvements in fit, discrimination, and overall cost allocation (>\$3 M reduction in error per 10,000 people).

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12889-020-08735-0>.

Additional file 1: Appendix. Additional details on the administrative claims dataset, input predictors, machine learning models, linear regression models, and statistical analysis. **Table S1.** Definitions and Conceptual Justification of the SDH Indicators. **Table S2.** Demographic Statistics of the Data Subsets by Geographic Location. **Table S3.** Variable Importances of the Prospective Linear Regression, LASSO Regression, Random Forest, and LightGBM Models without SDH Indicators. **Table S4.** Variable Importances of the Prospective Linear Regression, LASSO Regression, Random Forest, and LightGBM Models with SDH Indicators. **Table S5.** Performance Measures of LASSO Regression, Random Forest, and Multilayer Perceptron on the Test Set. **Table S6.** Predictive Ratio and Net Compensation Values of Prospective Machine Learning Models on Age and Sex Subgroups in the Test Set. **Table S7.** Predictive Ratio and Net Compensation Values of Prospective Linear Models on SDH-Based Subgroups in the Test Set. **Table S8.** Predictive Ratio and Net Compensation Values of Prospective Linear Models on Age and Sex Subgroups in the Test Set. **Figure S1.** Binned Scatter Plots of the Prospective Linear Regression and Machine Learning Models without and with SDH Indicators on the Test Set. **Table S9.** Performance Measures of Models Derived Using Binary Diagnosis Predictors on the Test Set. **Table S10.** Performance Measures of Top-Coded and Non-Top-Coded Models on the Test Set. **Table S11.** Performance Measures of Models with Lab Results on the Test Set. **Table S12.** Performance Measures of Concurrent and Prospective Models with SDH Indicators on the Test Set.

Abbreviations

ML: Machine learning; SDH: Social determinants of health; ACS: American Community Survey; MAE: Mean absolute error; C-statistic: Concordance statistic; CI: Confidence interval

Acknowledgements

Not applicable.

Authors' contributions

Conceptualization: SB and JAI. Design: JAI, AAK, MK, BH, PR. Data analysis and interpretation: JAI, AAK, MK, BH. Drafting of the manuscript: JAI, SB, AAK, MK. Critical revision of the manuscript for important intellectual content: SB, PR, BEL, LP, SP, AYN. Supervision: SB and AYN. The authors read and approved the final manuscript.

Funding

None.

Availability of data and materials

The code used in this study is shared online for reproducibility at: <https://github.com/stanfordmlgroup/risk-adjustment-ml>.

Ethics approval and consent to participate

Ethics approval was obtained by the Stanford Institutional Review Board (eProtocol #42334) and the need for consent was also waived by that IRB.

Consent for publication

Not applicable.

Competing interests

Dr. Landon currently serves as a clinical consultant to RTI, inc., related to ongoing updating and improvement of the Medicare HCC models and models used on the ACA exchanges. Dr. Basu receives salary from Collective Health, Inc., which uses risk adjustment models to risk stratify healthcare claims data to advise self-insured employers on care management strategies. Dr. Basu has also received consulting fees from KPMG for advising on risk models used to identify individuals susceptible to uncontrolled diabetes. None of these entities were involved in the formulation of the research design, its conduct, review of results, or decision to submit the research for publication. None of the other authors have any competing interests.

Author details

¹Department of Computer Science, Stanford University, 353 Serra Mall, Stanford, CA 94305, USA. ²Department of Statistics, Stanford University, Stanford, USA. ³Department of Healthcare Policy, Harvard Medical School, Boston, USA. ⁴Center for Primary Care, Harvard Medical School, Boston, USA. ⁵Center for Professionalism & Value in Health Care, American Board of Family Medicine Foundation, Lexington, USA. ⁶Robert Graham Center, American Academy of Family Physicians, Leawood, USA. ⁷Research and Analytics, Collective Health, San Francisco, USA. ⁸School of Public Health, Imperial College London, London, England.

Received: 8 January 2020 Accepted: 20 April 2020

Published online: 01 May 2020

References

- McGuire TG, Kleef RCV. Risk adjustment, risk sharing and premium regulation in health insurance markets: theory and practice. 1st ed. Oxford: Academic; 2018. p. 648.
- Ellis RP, Martins B, Rose S. Chapter 3 - risk adjustment for health plan payment. In: TG MG, van Kleef RC, editors. Risk adjustment, risk sharing and premium regulation in health insurance markets: Academic; 2018. p. 55–104. [cited 2019 Mar 9]. Available from: <http://www.sciencedirect.com/science/article/pii/B9780128113257000038>.
- Hileman G, Steele S. Accuracy of claims-based risk scoring models: Society of Actuaries; 2016. Available from: <https://www.soa.org/research-reports/2016/2016-accuracy-claims-based-risk-scoring-models/>.
- Rose S. A machine learning framework for plan payment risk adjustment. Health Serv Res. 2016 Dec;51(6):2358–74.
- Kan HJ, Kharrazi H, Chang H-Y, Bodycombe D, Lemke K, Weiner JP. Exploring the use of machine learning for risk adjustment: a comparison of standard and penalized linear regression models in predicting health care costs in older adults. PLoS One. 2019 Mar 6;14(3):e0213258.
- Kautter J, Pope GC, Ingber M, Freeman S, Patterson L, Cohen M, et al. The HHS-HCC risk adjustment model for individual and small group markets under the affordable care act. Medicare Medicaid Res Rev. 2014;4(3) [cited 2019 Nov 27]. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4214270/>.
- Park S, Basu A. Alternative evaluation metrics for risk adjustment methods. Health Econ. 2018;27(6):984–1010.
- Basu S, Berkowitz SA, Seligman H. The monthly cycle of hypoglycemia: an observational claims-based study of emergency room visits, hospital admissions, and costs in a commercially insured population. Med Care. 2017 Jul;55(7):639.
- Fitzpatrick T, Rosella LC, Calzavara A, Petch J, Pinto AD, Manson H, et al. Looking beyond income and education: socioeconomic status gradients among future high-cost users of health care. Am J Prev Med. 2015 Aug 1; 49(2):161–71.
- Field KS, Briggs DJ. Socio-economic and locational determinants of accessibility and utilization of primary health-care. Health Soc Care Community. 2001;9(5):294–308.
- Alley DE, Asomugha CN, Conway PH, Sanghavi DM. Accountable health communities — addressing social needs through Medicare and Medicaid. N Engl J Med. 2016 Jan 7;374(1):8–11.
- Whitehead M. Inequalities. In: Townsend P, Davidson N, Davidsen N, editors. Health: the black report/the health divide. London: Penguin UK; 1999. p. 464.
- A New Way to Talk about the Social Determinants of Health. 2010 RWJF. [cited 2020 Mar 17]. Available from: <https://www.rwjf.org/en/library/research/2010/01/a-new-way-to-talk-about-the-social-determinants-of-health.html>.
- Seligman HK, Laraia BA, Kushel MB. Food insecurity is associated with chronic disease among low-income NHANES participants. J Nutr. 2010 Feb; 140(2):304–10.
- Modrek S, Stuckler D, McKee M, Cullen MR, Basu S. A review of health consequences of recessions internationally and a synthesis of the US response during the great recession. Public Health Rev. 2013 Jun;35(1):1–33.
- 2012–2016 American Community Survey 5-year estimates. 2016. U.S. Census Bureau [cited 2019 Mar 11]. Available from: https://www.socialexplorer.com/data/ACS2016_5yr/metadata/?ds=ACS16_5yr.
- Ash AS, Mick EO, Ellis RP, Kiefe CI, Allison JJ, Clark MA. Social determinants of health in managed care payment formulas. JAMA Intern Med. 2017; 177(10):1424–30.
- Clinical Classifications Software (CCS) for ICD-10-PCS (beta version). [cited 2019 Sep 22]. Available from: <https://www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp>.
- Berkman LF, Kawachi I, Glymour MM. Social epidemiology. Oxford: Oxford University Press; 2014. p. 641.
- Closing the gap in a generation : health equity through action on the social determinants of health : Commission on Social Determinants of Health final report. - NLM Catalog - NCBI [Internet]. [cited 2019 Mar 29]. Available from: <https://www.ncbi.nlm.nih.gov/nlmcatalog/101488674>.
- Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction, second edition. 2nd ed. New York: Springer; 2016. p. 745.
- Ash AS, Ellis RP, Pope GC, Ayanian JZ, Bates DW, Burstin H, et al. Using diagnoses to describe populations and predict costs. Health Care Financ Rev. 2000;21(3):7–28.
- Pope GC, Kautter J, Ellis RP, Ash AS, Ayanian JZ, Iezzoni LI, et al. Risk adjustment of Medicare capitation payments using the CMS-HCC model. Health Care Financ Rev. 2004;25(4):119–41.
- Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
- Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, et al. Strong rules for discarding predictors in lasso-type problems. J R Stat Soc Series B Stat Methodology. 2012;74(2):245–66.
- Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–232.
- Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann Stat. 2000 Apr;28(2):337–407.
- Zhang H, Si S, Hsieh C-J. GPU acceleration for large-scale tree boosting; 2018.
- Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA. 2017;318(6):517–8.
- Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in neural

- information processing systems 30. Curran Associates, Inc.; 2017. p. 3146–3154. [cited 2019 Mar 11]. Available from: <http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf>.
31. Breiman L. Random forests. *Mach Lang*. 2001;45(1):5–32.
 32. Hinton GE. Connectionist learning procedures. *Artif Intell*. 1989;40(1): 185–234.
 33. Tibshirani R, Efron B. An introduction to the bootstrap. CRC Press; 1994. [cited 2018 Feb 22]. Available from: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.473.2742>.
 34. FEH J. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Corrected ed. New York: Springer; 2001. p. 572.
 35. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 1996;15(4):361–87.
 36. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. Hoboken: Wiley; 2000.
 37. Newson R. Confidence intervals for rank statistics: Somers' D and extensions. *Stata J*. 2006;6(3):309–34.
 38. Shepard M. Hospital network competition and adverse selection: evidence from the Massachusetts health insurance exchange: National Bureau of Economic Research; 2016. [cited 2020 Mar 22]. Report No.: 22600. Available from: <http://www.nber.org/papers/w22600>.
 39. Zink A, Rose S. Fair regression for health care spending. *ArXiv190110566 Cs Stat*. 2019; [cited 2019 Mar 18]; Available from: <http://arxiv.org/abs/1901.10566>.
 40. Pope GC, Kautter J, Ingber KJ, Freeman S, Sekar R, Newhart C. Evaluation of the CMS-HCC risk adjustment model, final report; 2011. p. 127.
 41. Bergquist SL, Layton TJ, McGuire TG, Rose S. Intervening on the data to improve the performance of health plan payment methods: National Bureau of Economic Research; 2018. [cited 2019 Mar 18]. Report No.: 24491. Available from: <http://www.nber.org/papers/w24491>.
 42. Hoef JMV. Who invented the Delta method? *Am Stat*. 2012;66(2):124–7.
 43. Rossum G. Python reference manual. Amsterdam: CWI (Centre for Mathematics and Computer Science); 1995.
 44. R Core Team. R: a language and environment for statistical computing: R Foundation for Statistical Computing; 2019. [cited 2019 Feb 11]. Available from: <https://www.R-project.org/>.
 45. Durfey SNM, Kind AJH, Gutman R, Monteiro K, Buckingham WR, DuGoff EH, et al. Impact of risk adjustment for socioeconomic status on Medicare advantage plan quality rankings. *Health Aff (Millwood)*. 2018;37(7):1065–72.
 46. Maddox KEJ, Reidhead M, Hu J, Kind AJH, Zaslavsky AM, Nagasaki EM, et al. Adjusting for social risk factors impacts performance and penalties in the hospital readmissions reduction program. *Health Serv Res*. 2019; 54(2):327–36.
 47. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019 Oct 25; 366(6464):447–53.
 48. Funk MJ, Landi SN. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. *Curr Epidemiol Rep*. 2014;1(4):175–85.
 49. Moore J, Hall J. The validity of claims-based risk estimation in underinsured populations. *Am J Manag Care*. 2012;18:e468–76.
 50. Wagner TH, Almenoff P, Francis J, Jacobs J, Chee CP. Assessment of the Medicare advantage risk adjustment model for measuring veterans affairs hospital performance. *JAMA Netw Open*. 2018;1(8):e185993.
 51. López-De Fede A, Stewart JE, Hardin JW, Mayfield-Smith K. Comparison of small-area deprivation measures as predictors of chronic disease burden in a low-income population. *Int J Equity Health*. 2016;15 [cited 2019 May 2]. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4901405/>.
 52. Before Penalizing Hospitals, Consider SDOH. *NEJM catalyst*; 2016. [cited 2019 Dec 1]. Available from: <https://catalyst.nejm.org/penalizing-hospitals-account-social-determinants-of-health/>.
 53. Dudley RA. The best of both worlds? Potential of hybrid prospective/concurrent risk adjustment. *Med Care*. 2003;41(1):56–69.
 54. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: The MIT Press; 2016. p. 775.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

