Predicting Self-Rated Health Across the Life Course: Health Equity Insights from Machine Learning Models

Cheryl R. Clark, MD, ScD[1,2,3], Mark J. Ommerborn, MPH[1], Kaitlyn Moran, MPH[1], Katherine Brooks, MD[3], Jennifer Haas, MD[4], David W. Bates, MD, MS[3], and Adam Wright, PhD[3]

[1]Center for Community Health and Health Equity, Brigham and Women's Hospital, 1620 Tremont Street, Boston, MA 02120, Boston, MA, USA; [2]Harvard Medical School, Boston, MA, USA; [3]Division of General Medicine and Primary Care, Brigham and Women's-Faulkner Hospitalist Program, Boston, MA, USA; [4]Division of General Medicine and Primary Care, Massachusetts General Hospital, Boston, MA, USA.

BACKGROUND: Self-rated health is a strong predictor of mortality and morbidity. Machine learning techniques may provide insights into which of the multifaceted contributors to self-rated health are key drivers in diverse groups.

OBJECTIVE: We used machine learning algorithms to predict self-rated health in diverse groups in the Behavioral Risk Factor Surveillance System (BRFSS), to understand how machine learning algorithms might be used explicitly to examine drivers of self-rated health in diverse populations.

DESIGN: We applied three common machine learning algorithms to predict self-rated health in the 2017 BRFSS survey, stratified by age, race/ethnicity, and sex. We replicated our process in the 2016 BRFSS survey.

PARTICIPANTS: We analyzed data from 449,492 adult participants of the 2017 BRFSS survey.

MAIN MEASURES: We examined area under the curve (AUC) statistics to examine model fit within each group. We used traditional logistic regression to predict self-rated health associated with features identified by machine learning models.

KEY RESULTS: Each algorithm, regularized logistic regression (AUC: 0.81), random forest (AUC: 0.80), and support vector machine (AUC: 0.81), provided good model fit in the BRFSS. Predictors of self-rated health were similar by sex and race/ethnicity but differed by age. Socioeconomic features were prominent predictors of self-rated health in mid-life age groups. Income [OR: 1.70 (95% CI: 1.62–1.80)], education [OR: 2.02 (95% CI: 1.89, 2.16)], physical activity [OR: 1.52 (95% CI: 1.46–1.58)], depression [OR: 0.66 (95% CI: 0.63–0.68)], difficulty concentrating [OR: 0.62 (95% CI: 0.58–0.66)], and hypertension [OR: 0.59 (95% CI: 0.57–0.61)] all predicted the odds of excellent or very good self-rated health.

CONCLUSIONS: Our analysis of BRFSS data show social determinants of health are prominent predictors of self-rated health in mid-life. Our work may demonstrate promising practices for using machine learning to advance health equity.

Received February 28, 2020
Accepted December 10, 2020
Published online February 23, 2021

KEY WORDS: machine learning; social determinants of health; self-rated health; socioeconomic factors; healthcare disparities.

J Gen Intern Med 36(5):1181–8
DOI: 10.1007/s11606-020-06438-1

© Society of General Internal Medicine 2021

INTRODUCTION

Promoting self-rated health is important in primary care and population health settings.[1–3] Examining individuals' self-assessment of their health shifts our focus as clinicians from treating illnesses to promoting wellness.[4] Identifying contributors to patients' self-assessments of their health may promote health equity in care by helping clinicians prioritize factors that can be addressed to meet patients' concerns. Self-rated health is a commonly used patient self-assessment measure in clinical and public health data, and is also associated with several hard indicators of health status including health care costs, the presence and severity of chronic disease, and the risk of mortality.[3] Because key contributors to an individual's sense of self-rated health are varied, include social as well as medical factors, and may differ in diverse groups, achieving equity in self-rated health may require active efforts to identify factors that are important within specific subgroups.[3,5] However, clinical data rarely collect information related to social factors that may influence self-rated health, and strategies to integrate social and clinical data to better understand concerns in specific groups are lacking.[4]

As many features contribute to self-rated health, the process of identifying key associated factors may be suited to a big data approach that considers multiple interrelated dimensions that are predictive in diverse groups. Machine learning allows consideration of multiple exposures simultaneously and hypotheses are not prespecified.[6] Methods for applying machine learning algorithms to address health equity as a part of prediction have not been fully described.

We applied common machine learning algorithms to analyze data from the Behavioral Risk Factor Surveillance System (BRFSS), to gain insights from a large population-based data resource designed to measure several social, demographic, healthcare utilization, and behavioral factors that might contribute to self-rated health. We examined whether self-rated health could be predicted with good model fit in diverse groups, using features from the BRFSS. Formally, our study objectives were twofold, (1) to build predictive models of self-rated health using machine learning algorithms applied to the 2016 and 2017 BRFSS, and (2) examine the accuracy and insights gained from machine learning models predicting excellent or very good self-rated health in diverse groups, along the lines of sex, race/ethnicity, and in age groups across the life course in the US.

METHODS

Data Sources and Study Population: the Behavioral Risk Factor Surveillance System

We chose the BRFSS as a population-based data source due to the large sample size and sociodemographic data available for examining self-rated health in multiple subgroups. The BRFSS is a cross-sectional, random digit-dial telephone survey of the non-institutionalized civilian population aged 18 years and older in the United States (US). The survey is administered by the Centers for Disease and Control and Prevention (CDC), and fielded annually by state health departments in the 50 states, the District of Columbia, and select US territories.[5] Raking weights are used to produce population estimates that adjust for survey non-coverage, non-response, and the probability of being sampled given the geographic location, age, race, and sex of the participant.[7]

Analytic Sample. We included all participants in the 2017 BRFSS, the most recent data at the time of the analysis (N = 449,492) and repeated the analysis in the 2016 BRFSS sample (N = 484,964) to validate our approach. We anticipated that a legacy of systemic racism in the US, along with the known differences in self-rated health by age and sex, could contribute to differences in model fit, and potentially yield different predictors of self-rated health by race, ethnicity, sex, and age.[4] Thus, within each survey year, we analyzed data for the entire cohort, and stratified analyses along three dimensions, by (1) sex, (2) race/ethnicity, or (3) age category: 18–29 years old, 30–39 years old, 40–49 years old, 50–59 years old, 60–69 years old, and 70 years and older.

Study Outcomes

Target Feature. The primary outcome of interest was self-rated health, which was measured as, “Would you say that in general your health is: excellent, very good, good, fair, or poor.” We classified individuals into excellent or very good health, compared to good, fair, or poor health.

Feature Selection, Inclusion, and Exclusion Criteria

A detailed list of 2017 BRFSS features included and excluded from the analysis is presented in Appendix Figure 1. We excluded features related to the landline/cell phone survey sampling components that were age- or sex-specific (e.g., prostate-specific antigen screening, mammography), that were derivative of a feature already included in the analysis, or that were closely related to self-rated health (e.g., health-related quality of life, number of poor mental health days). Lastly, we excluded features that might be unreliable due to missing values (item non-response greater than 50% of population).

To focus our analysis, we used a conceptual model—the Healthy People 2020 framework—to categorize the remaining 51 features for model inclusion that defined seven domains: *demographics*, which included age, sex, race, geographic division, state of residence, number of adults in the respondent’s household, marriage status, veteran status, number of children, and language spoken; *clinical conditions*, which included a self-reported history of cancer, asthma, depression, diabetes, stroke, cardiovascular disease, kidney disease, arthritis, COPD, skin cancer, body mass index, angina, or hypertension; *functional status*, which included difficulty doing errands, difficulty dressing, difficulty walking, difficulty communicating, blindness or deafness; *access to clinical care*, which included delayed care due to cost, having a primary care physician, insurance status, and having had doctor visit in the previous year; *health behavior*, which included alcohol use, smoking status, e-cigarette use, use of chewing tobacco, exercise practices, drunk driving, seat belt use, Internet use in last 30 days, daily fruit consumption, and daily vegetable consumption; *preventive care*, which included having had an HIV test, having identified HIV risk factors, and having had a flu vaccine; and *socioeconomic status*, which included education attainment, income category, homeownership, employment, and cell phone use.[8] To account for differences in scale, all features were formatted as binary dummy variables for analysis.

Data Analysis

Descriptive Data. We presented percentages for the top features of importance by age group (Table 1). Feature importance was determined by machine learning classification described below.

Machine Learning in R. We compared predictions and model fit for three supervised machine learning algorithms applied to 2017 BRFSS data to identify features predictive of “excellent” or “very good” self-rated health, compared to good, fair, or poor health. To build each model, we compared three algorithms, regularized logistic regression, random forest, and support vector machine algorithms in the Caret package of R software, version 3.4.0, using a high-performance computer cluster.[9] We split the data into two-thirds training data and one-third testing data. We used the bootstrapping resampling method during model training, which selects a random sample of the population “with replacement,” so that the full population is resampled with each model iteration. To examine model accuracy for diverse groups, we estimated model fit

Table 1 2017 BRFSS Descriptive Data by Age Groups

	Total *N* = 449,492	18–29 *N* = 48,649	30–39 *N* = 51,313	40–49 *N* = 56,931	50–59 *N* = 74,809	60–69 *N* = 92,844	70+ *N* = 106,199
Self-rated health							
Excellent or very good	220,176 (49.0)	29,274 (60.2)	29,019 (56.6)	29,579 (52.0)	35,855 (47.9)	43,105 (46.4)	44,613 (42.0)
Good, fair, or poor	229,316 (51.0)	19,375 (39.8)	22,294 (43.4)	27,352 (48.0)	38,954 (52.1)	49,739 (53.6)	61,586 (58.0)
Race							
Non-Hispanic White	374,432 (83.3)	34,513 (70.9)	33,376 (65.0)	43,960 (77.2)	60,808 (81.3)	80,249 (86.4)	93,902 (88.4)
Non-Hispanic Black	36,839 (8.2)	5106 (10.5)	8065 (15.7)	5771 (10.1)	6689 (8.9)	8502 (9.2)	8064 (7.6)
Hispanic	38,221 (8.5)	9030 (18.6)	8021 (15.6)	7200 (12.7)	8094 (10.8)	5107 (5.5)	3996 (3.8)
Sex							
Female	250,823 (55.8)	25,353 (52.1)	27,059 (52.7)	30,803 (54.1)	41,600 (55.6)	52,384 (56.4)	65,168 (61.4)
Male	198,669 (44.2)	23,296 (47.9)	24,254 (47.3)	26,128 (45.9)	33,209 (44.4)	40,460 (43.6)	41,031 (38.6)
Smoking status							
Current	69,100 (15.4)	8089 (16.6)	10,974 (21.4)	10,700 (18.8)	14,495 (19.4)	13,532 (14.6)	8111 (7.6)
Former	123,884 (27.6)	4794 (9.9)	10,132 (19.8)	11,727 (20.6)	17,603 (23.5)	30,348 (32.7)	42,772 (40.3)
Never	256,508 (57.1)	35,766 (73.5)	30,207 (58.9)	34,504 (60.6)	42,711 (57.1)	48,964 (52.7)	55,316 (52.1)
Body mass index							
Normal BMI	151,734 (33.8)	24,266 (49.9)	17,865 (34.8)	16,946 (29.8)	21,673 (29.0)	27,270 (29.4)	38,319 (36.1)
Overweight	157,877 (35.1)	13,856 (28.5)	17,113 (33.4)	19,578 (34.4)	26,359 (35.2)	33,646 (36.2)	40,644 (38.3)
Obese	139,881 (31.1)	10,527 (21.6)	16,335 (31.8)	20,407 (35.9)	26,777 (35.8)	31,928 (34.4)	27,236 (25.7)
Hypertension							
Yes	181,414 (40.4)	4523 (9.3)	8443 (16.5)	15,096 (26.5)	29,526 (39.5)	47,989 (51.7)	66,230 (62.4)
No	268,078 (59.6)	44,126 (90.7)	42,870 (83.6)	41,835 (73.5)	45,283 (60.5)	44,855 (48.3)	39,969 (37.6)
Depression							
Yes	89,662 (20.0)	10,227 (21.0)	11,048 (21.5)	12,361 (21.7)	17,335 (23.2)	19,919 (21.4)	15,004 (14.1)
No	359,830 (80.1)	38,422 (79.0)	40,265 (78.5)	44,570 (78.3)	57,474 (76.8)	72,925 (78.6)	91,195 (85.9)
Difficulty concentrating							
Yes	47,101 (10.5)	6217 (12.8)	5226 (10.2)	6149 (10.8)	11,405 (15.3)	9426 (10.2)	9801 (9.2)
No	402,391 (89.5)	42,432 (87.2)	46,087 (89.8)	50,782 (89.2)	63,404 (84.8)	83,418 (89.9)	96,398 (90.8)
Internet use							
Yes	366,415 (81.5)	46,967 (96.5)	48,499 (94.5)	52,314 (91.9)	64,126 (85.7)	74,505 (80.3)	64,639 (60.9)
No	83,077 (18.5)	1682 (3.5)	2814 (5.5)	4617 (8.1)	10,683 (14.3)	18,339 (19.8)	41,560 (39.1)
Education							
Less than high school	33,062 (7.4)	3230 (6.6)	3908 (7.6)	4192 (7.4)	5452 (7.3)	5576 (6.0)	9489 (8.9)
High school graduate	122,546 (27.3)	15,563 (32.0)	11,773 (22.9)	12,676 (22.3)	20,563 (27.5)	24,005 (25.9)	32,866 (31.0)
Some college	124,794 (27.8)	16,262 (33.4)	13,736 (26.8)	14,804 (26.0)	20,524 (27.4)	26,648 (28.7)	27,605 (26.0)
College graduate	169,090 (37.6)	13,594 (28.0)	21,896 (42.7)	25,259 (44.4)	28,270 (37.8)	36,615 (39.4)	36,239 (34.1)

Notes: Data imputed for missing values. Percentages unweighted and represent the characteristics of survey respondents but not population estimates

and performance using parameters on accuracy, area under the curve (AUC), and receiver operator curves (ROC) for the entire population and within subgroups. An AUC value provides a summary of model prediction visualized through ROC curves. A perfect model would have an AUC of 1.0 (perfect prediction of self-rated health), while a random model would have an AUC of 0.5 (chance prediction of self-rated health). Each algorithm used the ROC to identify a machine learning "importance factor," which is the rank list of the top twenty factors that contributed to prediction strength in the model. We examined the number of features identified within each of the seven domains and counted the number of times each feature was identified as one of the top twenty factors contributing to model fit.

Multiple Imputation for Machine Learning Analysis in R. To prevent bias from list-wise deletion in the analyzed data, we imputed missing values using multiple imputation techniques from the MICE (Multivariate Imputations by Chained Equations) package in R.[10] The MICE algorithm used a predictive mean matching technique to impute missing values using logistic regression.

The largest sources of missing data were due to missing responses for income (16.7%), HIV testing (12.2%), and daily vegetable intake (10.5%). To determine differences due to multiple imputation of missing data, we compared frequencies for each feature for imputed vs. non-imputed data. The absolute percentage difference between imputed and non-imputed values was between 0.01% (e-cigarette use) and 6.9% (non-Hispanic White race/ethnicity).

Odds of Excellent or Very Good Health. To improve interpretation of models, we used logistic regression to estimate the odds of excellent or very good health compared to good, fair, or poor health. We used the top 20 features identified by machine learning models to fit weighted logistic regression models for each population subgroup, accounting for the complex survey design in SAS via the SURVEYLOGISTIC procedure. To prevent bias from listwise deletion, we imputed data in SAS via the MI/MIANALYZE procedure.[11,12]

Comparison to 2016 Data. Survey questions for the BRFSS differ by year. We repeated our process in the BRFSS 2016 survey to understand how our results differed when different covariates were introduced.

RESULTS

Sex, Race/Ethnicity, and Age Trends in Self-Rated Health and Other Descriptive Covariate Features

The weighted prevalence of self-rated health and other study covariates in the 2017 BRFSS is presented in Table 1. Of the 449,492 participants, nearly half rated their health as excellent or very good (49.0%). Most identified their race as non-Hispanic White (83.3%); a majority identified as female (55.8%). The percentage of the population who rated their health as excellent or very good decreased with age (60.2% 18–29 years old, 56.6% 30–39 years old, 52.0% 40–49 years old, 47.9% 50–59 years old, 46.4% 60–69 years old, and 42.0% 70 years and older, *p* value < 0.001).

The most notable differences in covariate features were seen by age group. The youngest population (18–29 years old) had the highest percentage of any physical activity (77.7%), normal BMI (49.9%), Internet use in the past 30 days (96.5%), low income (29.2%), and the lowest percentage of arthritis (4.6%), diabetes (2.1%), hypertension (9.3%), difficulty walking (3.0%), difficulty doing errands (4.3%), and being married (20.7%). The population 70 of years of age and older had the highest percentage with arthritis (52.7%), diabetes (22.4%), and hypertension (62.4%), and have difficulty walking (28.2%). The population aged 70 and older had the lowest percentage of current smokers (7.6%), self-reported depression (14.1%), and Internet use in the past 30 days (60.9%).

Machine Learning Model Fit

We calculated AUC values for the three machine learning algorithms for the total population (Appendix Table 2). The AUC values were nearly identical for the total population for regularized logistic regression (0.81), random forest (0.80), and support vector machine (0.81). Model fit and importance ranking were similar by race/ethnicity (non-Hispanic Black AUC 0.78, Hispanic AUC 0.79, non-Hispanic White AUC 0.81), and sex (female AUC 0.82, male AUC 0.79), but differed by age. We thus focused our data presentation on results related to age groups across the life course. For all three algorithms, AUCs were highest (0.79–0.83) among mid-life age groups (ages 40–49, 50–59 years) and young elders (ages 60–69 years). Model fit (AUCs 0.72–0.73) was slightly lower in the youngest age category (18–29 years). Due to similar fit across algorithms, we used regularized logistic regression to select importance features due to ease of interpretation.

Top Features Predicting Self-Rated Health from Machine Learning Algorithms

The top importance features identified by the regularized logistic regression machine learning algorithm across all age groups are presented in Figure 1. The top two features predicting excellent and very good health for the young population (18–29 years old) were BMI and depression. Education and income were the top features predicting self-rated health in the populations aged 30–39 and 40–49 years. The top two features predicting self-rated health in the population aged 50–59 were difficulty walking and income; for the age 60–69 population, the top two features were difficulty walking and hypertension; and in those aged 70 and older, the top two features were difficulty walking and Internet use in the past 30 days.

We counted the number of features identified within each covariate domain and plotted the number in Figure 2. For all groups, socioeconomic status features were important predictors of self-rated health across the life course and were identified most frequently by the regularized logistic regression machine learning model as predictors of self-rated health in mid-life (in the 30–39, 40–49, and 50–59 years age groups). Health behaviors and health care access were identified as important predictors of self-rated health for the population aged 18–29 years old. Comorbidities and functional status were important predictors of self-rated health for those aged 70 and older. Race/ethnicity was most frequently identified as a predictor of self-rated health in younger age groups (18–29, 30–39).

Odds of Excellent or Very Good Health in Frequentist Logistic Regression Models

We estimated the odds of excellent or very good self-rated health by age group, using the top 20 features that were important to predicting self-rated health, as identified by the regularized logistic regression machine learning algorithm (Table 2). The frequentist logistic models reduced to 20 importance features demonstrated good model fit (AUC values 0.72–0.82).

Higher income and education increased the odds of excellent or very good self-rated health in all age groups (Table 2). Additionally, physical activity, self-reported depression, having difficulty concentrating, and the presence of hypertension all predicted the odds of excellent or very good self-rated

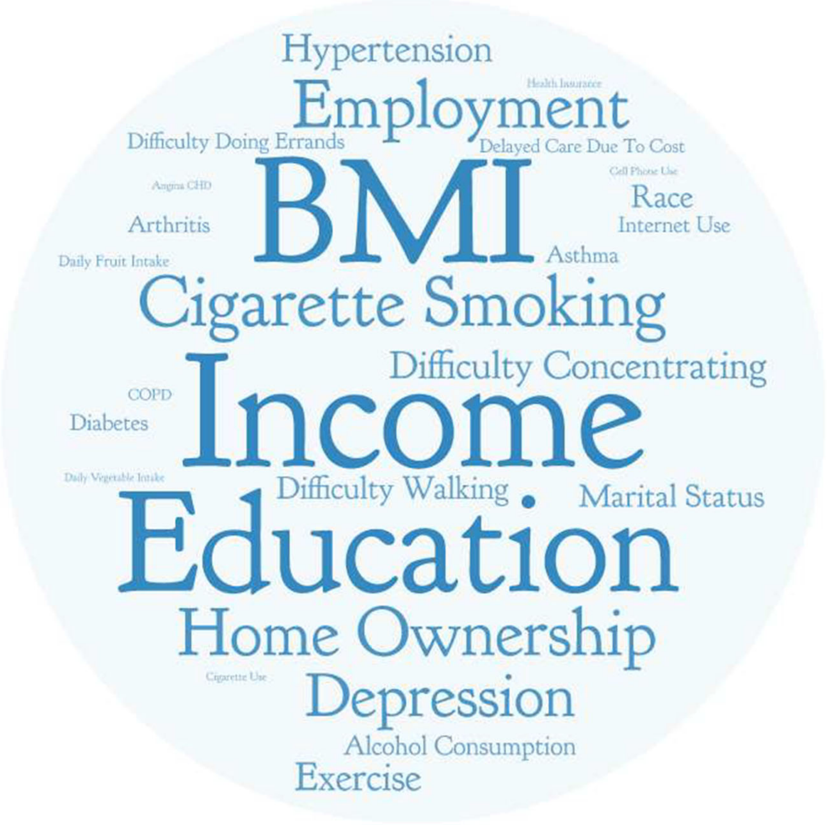

Figure 1 Top variables of importance across age groups, 2017 BRFSS. Notes: Seven domains include *demographics*: age, sex, race, geographic division, state of residence, number of adults in the respondent's household, marriage status, veteran status, number of children, and language spoken; *clinical conditions*: a self-reported history of cancer, asthma, depression, diabetes, stroke, cardiovascular disease, kidney disease, arthritis, COPD, skin cancer, body mass index, angina, or hypertension; *functional status*: difficulty doing errands, difficulty dressing, difficulty walking, difficulty communicating, blindness, or deafness; *access to clinical care*: delayed care due to cost, having a primary care physician, insurance status, and having had doctor visit in the previous year; *health behavior*: alcohol use, smoking status, e-cigarette use, use of chewing tobacco, exercise practices, drunk driving, seat belt use, Internet use in last 30 days, daily fruit consumption, and daily vegetable consumption; *preventive care*: having had an HIV test, having identified HIV risk factors, and having had a flu vaccine; *socioeconomic status*: education attainment, income category, homeownership, employment, and cell phone use.

health in all age groups. Increasing BMI was associated with decreasing odds of self-rated health in all groups except the oldest BRFSS participants (70+). Non-Hispanic Black race was associated with lower self-rated health in mid-life and older groups (age 50–59, age 60–69, age 70+). Sex was not identified as a predictive feature in any model. Full results for

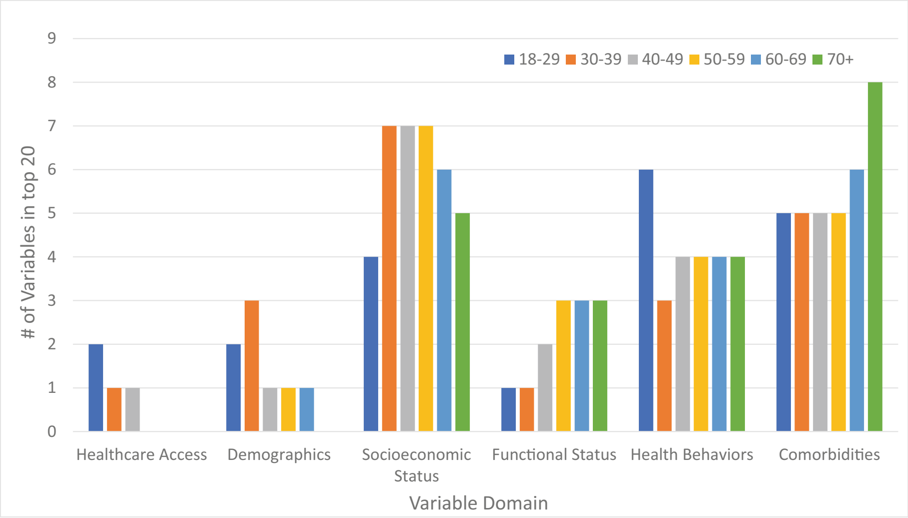

Figure 2 Top variables of importance by domain and age group, 2017 BRFSS. Notes: Data analysis performed with regularized logistic regression machine learning algorithm.

Table 2 2017 BRFSS Weighted Odds Ratios (95% CI) of Excellent or Very Good Self-Rated Health by Age Groups

	Total	**18-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70+**
	***N* = 449,492**	***N* = 48,649**	***N* = 51,313**	***N* = 56,931**	***N* = 74,809**	***N* = 92,844**	***N* = 106,199**
	AUC = 0.7928	**AUC = 0.7183**	**AUC = 0.7657**	**AUC = 0.7942**	**AUC = 0.8244**	**AUC = 0.8213**	**AUC = 0.7809**
Race							
Non-Hispanic White	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Non-Hispanic Black	0.84	0.94	1.03	1.05	0.84	0.69	0.57
	(0.80, 0.89)	(0.84, 1.06)	(0.91, 1.18)	(0.92, 1.19)	(0.74, 0.96)	(0.59, 0.80)	(0.49, 0.66)
Hispanic	0.60	0.67	0.63	0.66	0.51	0.58	0.51
	(0.57, 0.63)	(0.61, 0.75)	(0.56, 0.71)	(0.59, 0.75)	(0.44, 0.60)	(0.48, 0.68)	(0.42, 0.63)
Physical activity							
No activity	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Any activity	1.52	1.61	1.53	1.56	1.46	1.56	1.31
	(1.46, 1.58)	(1.46, 1.78)	(1.39, 1.67)	(1.41, 1.72)	(1.34, 1.60)	(1.43, 1.71)	(1.21, 1.43)
Body mass index							
Normal BMI	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Overweight	0.83	0.79	0.78	0.80	0.82	0.86	0.97
	(0.80, 0.86)	(0.72, 0.86)	(0.71, 0.85)	(0.72, 0.89)	(0.75, 0.89)	(0.79, 0.94)	(0.90, 1.05)
Obese	0.55	0.46	0.45	0.49	0.55	0.63	0.91
	(0.52, 0.57)	(0.42, 0.51)	(0.40, 0.50)	(0.44, 0.55)	(0.50, 0.60)	(0.58, 0.70)	(0.83, 1.00)
Depression							
Yes	0.66	0.62	0.65	0.69	0.64	0.74	0.72
	(0.63, 0.68)	(0.56, 0.68)	(0.59, 0.72)	(0.61, 0.77)	(0.58, 0.71)	(0.67, 0.81)	(0.64, 0.82)
No	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Difficulty concentrating							
Yes	0.62	0.61	0.58	0.68	0.57	0.58	0.64
	(0.58, 0.66)	(0.54, 0.69)	(0.49, 0.68)	(0.54, 0.85)	(0.48, 0.66)	(0.49, 0.70)	(0.56, 0.74)
No	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Income							
< $25,000	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$25–$49,999	1.13	1.16	1.26	1.10	1.06	1.12	1.05
	(1.08, 1.19)	(1.04, 1.29)	(1.12, 1.42)	(0.95, 1.27)	(0.92, 1.22)	(1.01, 1.24)	(0.95, 1.16)
$50–$74,999	1.35	1.29	1.46	1.42	1.29	1.29	1.21 (1.09, 1.34)
	(1.27, 1.43)	(1.14, 1.47)	(1.27, 1.69)	(1.21, 1.66)	(1.12, 1.48)	(1.16, 1.45)	
≥ $75,000	1.70	1.53	1.75	1.83	1.69	1.72	1.45
	(1.62, 1.80)	(1.37, 1.70)	(1.52, 2.02)	(1.59, 2.11)	(1.46, 1.96)	(1.53, 1.92)	(1.30, 1.61)
Education							
Less than high school	1.00	1.00	1.00	1.00	1.00	1.00	1.00
High school graduate	1.44	1.42 (1.23, 1.64)	1.65	1.55	1.26	1.18	1.32
	(1.35, 1.54)		(1.41, 1.92)	(1.28, 1.87)	(1.07, 1.49)	(0.98, 1.41)	(1.16, 1.51)
Some college	1.69	1.61	2.12	1.73	1.49	1.45	1.61
	(1.58, 1.81)	(1.40, 1.87)	(1.81, 2.49)	(1.43, 2.09)	(1.26, 1.77)	(1.20, 1.75)	(1.40, 1.85)
College graduate	2.02	2.03	2.44	2.26	1.84	1.68	1.67
	(1.89, 2.16)	(1.74, 2.36)	(2.09, 2.86)	(1.86, 2.75)	(1.55, 2.20)	(1.39, 2.03)	(1.45, 1.93)
Hypertension							
Yes	0.59	0.63	0.57	0.51	0.56	0.57	0.64
	(0.57, 0.61)	(0.55, 0.71)	(0.51, 0.63)	(0.47, 0.56)	(0.52, 0.60)	(0.54, 0.62)	(0.60, 0.69)
No	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Notes: Weighted analysis performed via SurveyLogistic procedure in SAS. Data imputed for missing values via PROC MI/MIANALYZE

weighted logistic regression models by age groups are presented in Appendix Table 3.

Replication in 2016 Data

In the 2016 BRFSS data, patterns for model fit, importance ranking, and prediction in logistic regression were substantially like the 2017 data except for the few features which were only captured in specific years. For instance, dental health factors (last time you saw a dentist and number of teeth extracted) were ranked as important predictors of self-rated health, and this was surveyed in 2016 but not 2017.

DISCUSSION

Machine learning models are increasingly used to gain population health insights, though explicitly training models to provide insights on factors that matter for health in diverse populations has not been commonly reported. In the BRFSS, we found three common machine learning algorithms provided good model fit and predicted similar features as contributors to excellent and very good self-related health. Notably, we anticipated that racial and ethnic differences may have influenced how well algorithms predicted self-rated health, as well as the types of features that were important predictors of self-rated health in specific racial and ethnic groups. However, in these analyses, we found good model fit and predictions using these approaches that were similar by race/ethnicity and by sex. Instead, we found the factors that predicted self-rated health differed across the life course, where diseases of aging were clearly prominent in older groups, access to care and health behaviors were most predictive of model fit among the young, and socioeconomic indicators predicted self-rated health in all groups, but particularly in mid-life. In regression models that took advantage of the data reduction and refinement suggested by machine learning models, we were able to predict self-rated health with a moderately high degree of accuracy in most groups studied.

Our analysis of BRFSS data highlighted the importance of socioeconomic conditions in mid-life. The influence of socioeconomic status on health and well-being in mid-life has been previously recognized in the life course literature.[13] The cumulative effects of social conditions may become apparent in mid-life, while access to Medicare and survivorship bias potentially explains the decline in the impact of socioeconomic status in older age.[14]

Though the dimensions measured in the BRFSS are interrelated, visualizing the association of specific risks at different points across the life course reinforces the idea that a “precision health” approach is needed to influence self-rated health and well-being at specific stages of the life course, tailored and matched to population need.

We were able to identify the importance of socioeconomic factors to self-rated health due to their availability in public health data. Our findings underscore the importance of collecting socioeconomic data and asking patients about socioeconomic factors in clinical settings for use in electronic health records (EHRs), to make this information available in patient care settings for algorithmic prediction, and to plan for life-stage appropriate interventions that promote wellness for patients, including community partnerships that address social determinants of health.[15]

Limitations

A critical limitation of all machine learning models is that models capture associations that are not causal. For example, though we identified BMI and depression—alongside a series of factors—as predictors of self-rated health in young adults, it is not clear that intervening on these factors alone or in combination would influence self-rated health, or whether there is reverse causation, where those with poor self-rated health go on to develop higher BMI and depression. Thus, insights from machine learning models may be helpful for identifying relationships within data, but additional strategies, including implementation studies and other methods, are needed to develop actionable strategies for self-rated health interventions.

Second, our models were inherently limited by features available in BRFSS data. We noted that dental health predicted self-rated health in the 2016 data but were not available in the 2017 data. The limitations of data availability make it useful to employ a conceptual model to clearly outline the factors that are present or missing in models to facilitate data interpretation. For this purpose, we have chosen the Healthy People 2020 framework to identify relevant dimensions for analysis. Using this framework, we recognize many factors such as neighborhood and built-environment (segregation, area-level poverty, or deprivation), social and community context (social relationships and social capital, experiences of discrimination), detailed clinical data, and other factors not readily linked to individual-data surveyed in the BRFSS may affect the conclusions drawn here. One strength of our findings is the high AUC values suggesting good model fit with data that are available in the BRFSS. Additionally, though a strength of the BRFSS is the availability of rich survey data, including social and health status features, the BRFSS population represents those who made themselves available for telephone interviews, and may not fully represent the US population. Replicating these strategies in population health data may provide information that is relevant to clinical populations.

Importantly, using machine learning models as explicit tools to examine potential contributors of health inequities along the lines of race/ethnicity, sex, and other dimensions is a new area of research, and a strength of our current analysis.[16–18] The key strategies that we employed here: (1) using data that captured socioeconomic factors, and (2) stratifying the training and testing of machine learning algorithms in prespecified groups, we were able to confirm that models performed similarly by race/ethnicity and sex. We were also able to identify life course patterns that deserve further study and potential intervention. Future studies should also examine more detailed data on sexual orientation and gender identity (SOGI), which are becoming increasingly available in public datasets.

CONCLUSIONS

Using machine learning models in population-based data, we identified 20 out of 51 factors that are especially predictive of self-rated health in specific groups across the life course, with a moderately high degree of model fit. Though the findings in this study show similar predictions by race, ethnicity and other demographics, the strategy we propose to confirm and validate the accuracy of predictions in diverse groups is important to test assumptions that models are relevant to subgroups. As machine learning models are increasingly used in clinical and population health settings, it is important for clinicians, researchers, and population health managers to become facile with the use of strategies to ensure equity in the application of these methods.[19]

The strategies used here may enhance equity in the use of machine learning models when applied to EHR data, including (1) ensuring data sources capture relevant social, demographic, and contextual data on which to base predictions, (2) using conceptual models to provide transparency on factors that contribute or are left out of predictions, (3) stratifying models within specified subgroups to monitor the accuracy of model predictions in subgroups, and (4) reporting data by subgroup to inform interventions that may be needed within specific groups. Future research should examine these strategies for use in clinical data to enhance equity in prediction for diverse populations.

Supplementary Information *The online version contains supplementary material available at https://doi.org/10.1007/s11606-020-06438-1.*

Corresponding Author: *Cheryl R. Clark, MD, ScD; Center for Community Health and Health Equity, Brigham and Women's Hospital, 1620 Tremont Street, Boston, MA 02120, Boston, MA, USA (e-mail: crclark@partners.org).*

Funding *This research was supported with institutional funding from Brigham and Women's Hospital.*

Compliance with Ethical Standards:

Conflict of Interest: *The authors declare that they do not have a conflict of interest.*

REFERENCES

1. **DeSalvo KB, Jones TM, Peabody J, et al.** Health care expenditure prediction with a single item, self-rated health measure. Med Care. 2009;47(4):440-7.
2. **Boscardin CK, Gonzales R, Bradley KL, et al.** Predicting cost of care using self-reported health status data. BMC Health Serv Res. 2015;15(1):406.
3. **Balkrishnan R, Anderson RT, Bowton D**. Self-reported health status predictors of healthcare services utilization and charges in elderly asthmatic patients. J Asthma. 2000;37(5):415-23.
4. **Zimmerman FJ, Anderson NW**. Trends in Health Equity in the United States by Race/Ethnicity, Sex, and Income, 1993-2017. JAMA Netw Open. 2019;2(6):e196386. https://doi.org/10.1001/jamanetworkopen.2019.6386.
5. **Johnston KJ, Joynt Maddox KE**. The role of social, cognitive, and functional risk factors in Medicare spending for dual and nondual enrollees. Health Aff (Millwood). 2019;38(4):569-76.
6. **Waljee AK, Higgins PD**. Machine Learning in Medicine: a Primer for Physicians. Am J Gastroenterol. 2010;105(6):1224-1226. https://doi.org/10.1038/ajg.2010.173
7. Methodologic changes in the Behavioral Risk Factor Surveillance System in 2011 and potential effects on prevalence estimates. MMWR Morb Mortal Wkly Rep. 2012;61(22):410-3.
8. U.S. Department of Health and Human Services. Healthy People 2020 Washington, DC: Office of Disease Prevention and Health Promotion; [Available from: https://www.healthypeople.gov/sites/default/files/HP2020Framework.pdf. Accessed 21 January 2021.
9. **Kuhn M.** caret: Classification and Regression Training. R package version 60-8020018.
10. **van Buuren S, Groothuis-Oudshoorn K**. mice: Multivariate imputation by chained equations in R. 2011. 2011;45(3):67.
11. **Rubin DB**. Multiple imputation after 18+ years. J Am Stat Assoc. 1996;91(434):473.
12. **Schafer JL**. Multiple imputation: a primer. Stat Methods Med Res. 1999;8(1):3-15.
13. **Robert SA, Cherepanov D, Palta M, et al.** Socioeconomic status and age variations in health-related quality of life: results from the National Health Measurement Study. J Gerontol B Psychol Sci Soc Sci. 2009;64B(3):378-89.
14. **Herd P, Robert SA, House JS**. Health disparities among older adults. Elsevier; 2011. p. 121-34.
15. **Daniel H**, **Bornstein SS**, **Kane GC**. Addressing social determinants to improve patient care and promote health equity: an American College of Physicians position paper. Ann Intern Med. 168.8 (2018): 577-578.
16. **Panch T, Pearson-Stuttard J, Greaves F, et al.** Artificial intelligence: opportunities and risks for public health. Lancet Digit Health. 2019;1(1):e13-e4.
17. **Rajkomar A, Hardt M, Howell MD, et al.** Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866-72.
18. **Veinot TC, Ancker JS, Bakken S**. Health informatics and health equity: improving our reach and impact. JAMIA. 2019;26(8-9):689-95.
19. **Lin SY, Mahoney MR, Sinsky CA**. Ten ways artificial intelligence will transform primary care. J Gen Intern Med. 2019;34(8):1626-30.

Publisher's Note: *Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.*