PSYCHOLOGY, HEALTH & MEDICINE
2025, VOL. 30, NO. 6, 1158–1170
https://doi.org/10.1080/13548506.2025.2450546

The role of psychological factors in predicting self-rated health: implications from machine learning models

Jeong Ha (Steph) Choi[a*] and Daniel Hong Jung[b]

[a]Department of Psychology, Georgia State University, Atlanta, GA, USA; [b]Department of Public Policy and Management, University of Georgia, Athens, GA, USA

ABSTRACT

Self-rated health (SRH) is a significant predictor of future health outcomes. Despite the contribution of psychological factors in individuals' subjective health assessments, prior studies of machine learning-based prediction models primarily focused on health-related factors of SRH. Using the Midlife in the United States (MIDUS 2), the current study employed machine learning techniques to predict SRH based on a broad array of biological, psychological, and sociodemographic factors. Our analysis, involving logistic regression, LASSO regression, random forest, and XGBoost models, revealed robust predictive performance (AUPRC > 0.90) across all models. Emotion-related variables consistently emerged as vital predictors alongside health-related factors. The models highlighted the significance of psychological well-being, personality traits, and emotional states in determining individuals' subjective health ratings. Incorporating psychological factors into SRH prediction models offers a multifaceted perspective, enhancing our understanding of the complexities behind self-assessed health. This study underscores the necessity of considering emotional well-being alongside physical conditions in assessing and improving individuals' subjective health perceptions. Such insights hold promise for targeted interventions aimed at enhancing both physical health and emotional well-being to ameliorate subjective health assessments and potentially long-term health outcomes.

ARTICLE HISTORY
Received 16 February 2024
Accepted 31 December 2024

KEYWORDS
Self-rated health; machine learning; emotion; well-being

Introduction

Self-rated health, a global assessment of one's health, is an important proxy of future health such as morbidity and mortality, showing strong predictive power over and beyond other health risk factors (Benyamini & Idler, 1999; Franks et al., 2003; Idler et al., 2000; Latham & Peek, 2013; Wu et al., 2013; Wuorela et al., 2020). Largely operationalized as a measure of health-related quality of life (Fayers & Sprangers, 2002), its subjective nature complements clinical measures by considering the

CONTACT Jeong Ha (Steph) Choi jchoi118@gsu.edu Department of Psychology, Georgia State University, 140 Decatur Street, Suite 754 Urban Life Building, Atlanta, GA 30303, USA

*Jeong Ha (Steph) Choi served as lead for conceptualization, data curation, formal analysis, project administration. Daniel Hong Jung was lead in supervision and validation. Both authors equally contributed to methodology, software, visualization, writing – original draft, and writing – review & editing.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/13548506.2025.2450546

individual's overall well-being. Its practicality has thus proved beneficial for understanding individuals' health status (Jylhä et al., 1998; OECD, 2023) and improving self-rated health has been an important goal in clinical and population health settings (Boscardin et al., 2015; Desalvo et al., 2009). A large body of literature on self-assessed health and its correlates uncovered a variety of factors from general physical function and health behaviors (Krause & Jay, 1994; Molarius & Janson, 2002) to psychosocial factors (Cornwell & Waite, 2009; Idema et al., 2020). Together, studies largely agree on the nuanced and complex concept of subjective health.

Recent studies have begun to incorporate advanced computational methods predicting self-rated health to investigate the main contributors of self-assessed health and their relative importance (Chen et al., 2023; Clark et al., 2021; Gumà-Lao & Arpino, 2023). Using large-scale or population-based data, these studies have shown that both health-related and sociodemographic factors show high importance in predicting self-rated health, underscoring the variation in key contributors of self-assessed health. Despite such advances and benefits, there is a lack of consideration for psychological measures in previous studies on prediction models for self-rated health despite theoretical and empirical implications. The unique contribution of emotion on self-rated health, for instance, has been well established (Segerstrom, 2014; Watson & Pennebaker, 1989). Psychological well-being (Ryff et al., 2015), personal mastery, self-esteem (Cott et al., 1999), lower perceived stress (Svedberg et al., 2006), personality traits (Stephan et al., 2020), and regulatory efforts to achieve important health-related goals (Bailis et al., 2003) are all known to be associated with self-rated health. We thus argue that psychological factors in addition to health-related and environmental factors are important sources linked to self-assessed health and need to be considered when developing prediction models for self-rated health.

The present study therefore aimed to develop and compare prediction models via machine learning methods to predict self-rated health with a range of biological, psychological, and social determinants. We also calculated feature importance based on developed prediction models to understand the importance of each variable in predicting self-reported health. To achieve this aim, we used the Midlife in the Unites States (MIDUS) database, which includes data on not only health conditions and behavior, and socioenvironmental measures, but also a range of psychosocial measures.

Methods

Data source and study sample

We used the survey data from the second wave of Midlife in the United States Study (MIDUS 2; 2004 – 2006). MIDUS 2 is the first follow-up project of MIDUS 1, a nationally representative study of health and aging in the noninstitutionalized civilian population of the 48 contiguous United States (Brim et al., 2020; Ryff et al., 2021). MIDUS 2 also includes a city-specific oversample of African Americans to participate in a field interview and questionnaire that paralleled the main sample instruments (Radler, 2014). The data that support the findings of this study are openly available in ICPSR at https://doi.org/10.3886/ICPSR04652.v8. We restricted the analytic sample to those with available outcome data and less than 25% missing in the selected features, coming to a total of 4430 cases.

Feature selection and preprocessing

Features included in the model were selected if the feature was (1) present in both MIDUS RDD sample and Milwaukee sample, (2) not population-specific (e.g. pregnancy status), (3) about the respondent only, (4) not similar to or derived from other variables, (5) not similar to or overlap with self-rated health, and (6) provided with more than 70% of response from the entire sample. The final set came down to 394 features. Once features were selected, continuous variables were further scaled to range within 0 and 1 to ease the computational load. Categorical variables with three or more categories were also recoded into dichotomous variables. Full list of variables included is available in Appendix A.

Outcome

Self-rated health was our main predictor. Participants answered the following question: 'Using a scale from 0 to 10 where 0 means "the worst possible health" and 10 means "the best possible health", how would you rate your health these days?' Responses were categorized into two groups: 7 or higher (1), 6 or lower (0). Our reason for using a binary outcome was due to computational purposes (i.e. classification models are less likely to run into convergence issues) and based on prior studies (Chen et al., 2023; Clark et al., 2021; Gumà-Lao & Arpino, 2023).

Analytic Approach

Our study was conducted using Python 3.11, and all codes utilized for the research are accessible at https://osf.io/uva74/?view_only=5189c20c72c24c86bd102f253178e0e9. The analytical procedure included model development, assessing model performance and determining important features (for similar analytic approach, see Jung et al., 2023).

Missing data

To prevent bias from list-wise deletion in the analyzed data, we imputed missing values using the k nearest neighbor (kNN) imputation method (Emmanuel et al., 2021). The idea behind kNN imputation is to take advantage of positive correlations between cases. The assumption is that information about the missing values for a specific case is best provided by the *k* cases most similar to the case with the missing value. This approach identifies the neighboring points based on the calculated distance. The missing values are then estimated using the completed values of neighboring observations (the k nearest neighbors). This approach has been found effective in imputation outcomes compared to non-machine learning imputation approaches and less computationally taxing compared to random forest imputations (Emmanuel et al., 2021; Petrazzini et al., 2021). The number of nearest neighbors (k) was decided as part of the hyper-parameterization pipeline during model development.

Developing prediction models

The dataset was randomly divided into training and test sets (7:3 ratio) to avoid overfitting (Tan et al., 2018). We first constructed a traditional logistic regression model to predict self-rated health. We then developed three models (i.e. LASSO regression, random forest, and XGBoost) and compared them to the logistic regression model. Selection was based on their relative advantages, including transparency (Least Absolute Shrinkage and Selection Operator; LASSO; Balabaeva & Kovalchuk, 2021) and strong predictive performance (RF and XGBoost; Chang et al., 2019; Darabi et al., 2021; Taninaga et al., 2019). Least Absolute Shrinkage and Selection Operator (LASSO) regression is a penalized regression that employs the shrinkage method to prevent overfitting and facilitate feature selection (Göbl et al., 2015). Hyperparameter tuning, specifically for lambda, was conducted through 10-fold cross-validation. Random Forest (RF) is an ensemble method using bagged decision trees, and we optimized hyperparameters using grid search. eXtreme Gradient Boosting (XGBoost) is a decision tree algorithm designed for computational speed and performance (Liu et al., 2020). Hyperparameters including maximum tree depth, learning rate, subsample percentage, subsample ratio of variables, and maximum tree depth were tuned for this model.

Determining model performance

We evaluated the predictive performance of the models using ten-fold cross-validation, tuning hyperparameters for optimal average precision. The Area Under Curve for the Precision–Recall Curve (AUPRC) was utilized for performance comparison. AUPRC provides a single score summarizing model performance, where higher values indicate better performance (Boyd et al., 2013; Saito et al., 2015). AUPRC is considered informative for imbalanced data (Awan et al., 2019; Saito et al., 2015; Sofaer et al., 2019).

Examining feature importance

After comparing the performance of the developed models, we delved into the feature importance of the best-performing model. Feature importance indicates the values assigned to each ‘features’, or variables or factors, based on their relevance to the outcome (i.e. self-rated health). Higher scores indicate greater relevance to the outcome (Brick et al., 2017).

Results

Descriptive statistics

Table 1 presents the descriptive statistics of our sample.

Model performance

Figure 1 displays Area Under Precision–Recall Curves (AUPRCs) of the prediction models predicting self-rated health. All four models overall show high performance (AUPRC > 0.90). XGBoost showed the highest AUPRC (AUPRC = 0.946, confidence interval (CI): [0.935, 0.958]), followed by LASSO (AUPRC = 0.942, CI: [0.931, 0.954]), RF (AUPRC = 0.936, CI: [0.923, 0.950]), and logistic regression (AUPRC = 0.936, CI: [0.923, 0.948]). Following the high performance of the models, we examined the feature importance of both XGboost and RF models, as well as the included variables within the LASSO regression model.

Table 1. Descriptive statistics of final sample.

Variables	Statistics ($N = 4430$)
Age (M, SD)	55.88 (12.36)
Gender (Female %)	56.27
Education (%)	
High school or lower	35.03
Some college or 2-year college	38.99
4-year college or higher	35.98
Race (%)	
Non-Hispanic White	82.21
Non-Hispanic Black	11.83
Hispanic	2.71
Others	3.24
Marital Status (Married/Cohabitating %)	71.14
Employment Status (Employed %)	61.83

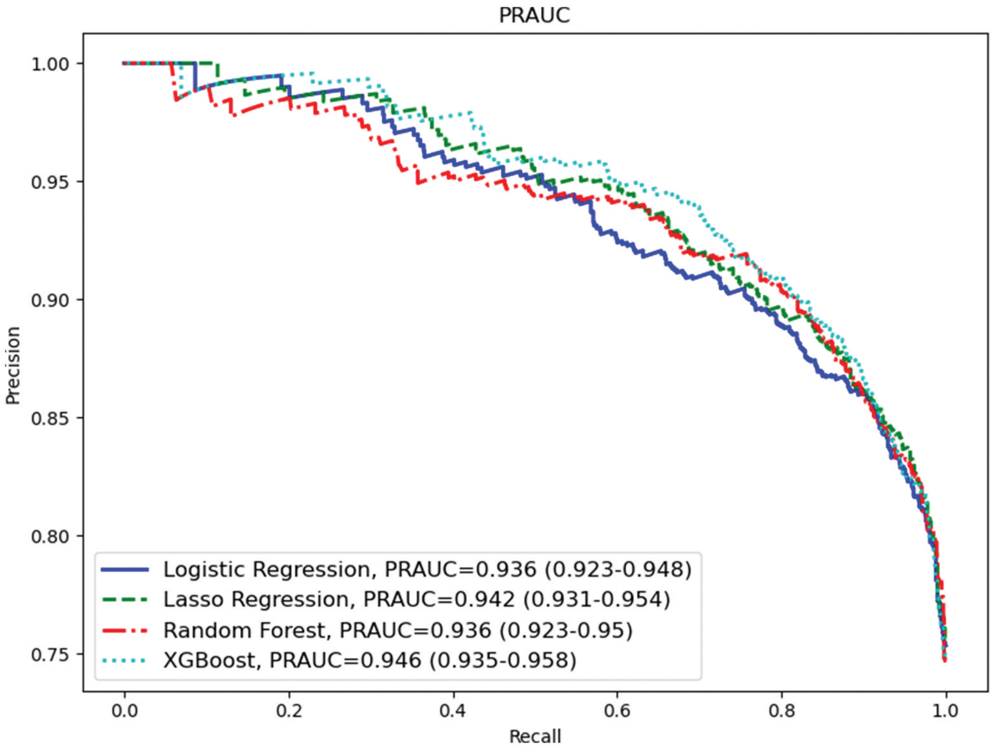

Figure 1. Precision-recall area under curve by prediction Model.

Feature importance

Figure 2 displays the feature importance of the model based on XGBoost and RF. Most items on functional limitations (e.g. health limits walking more than one mile, health limits climb one flight of stairs, health limits vigorous activity, health limits walking several blocks, health limits walking one block) and having diabetes or high blood sugar levels within the past year were within the top 20 features for both models. We also note that features that were not specific to health conditions or health behaviors present in the top list for both models. How positive (e.g. ‘felt full of life’, ‘good spirits’) or negative (e.g. ‘everything was an effort’) they felt in the past month were important features for both XGBoost and RF models.

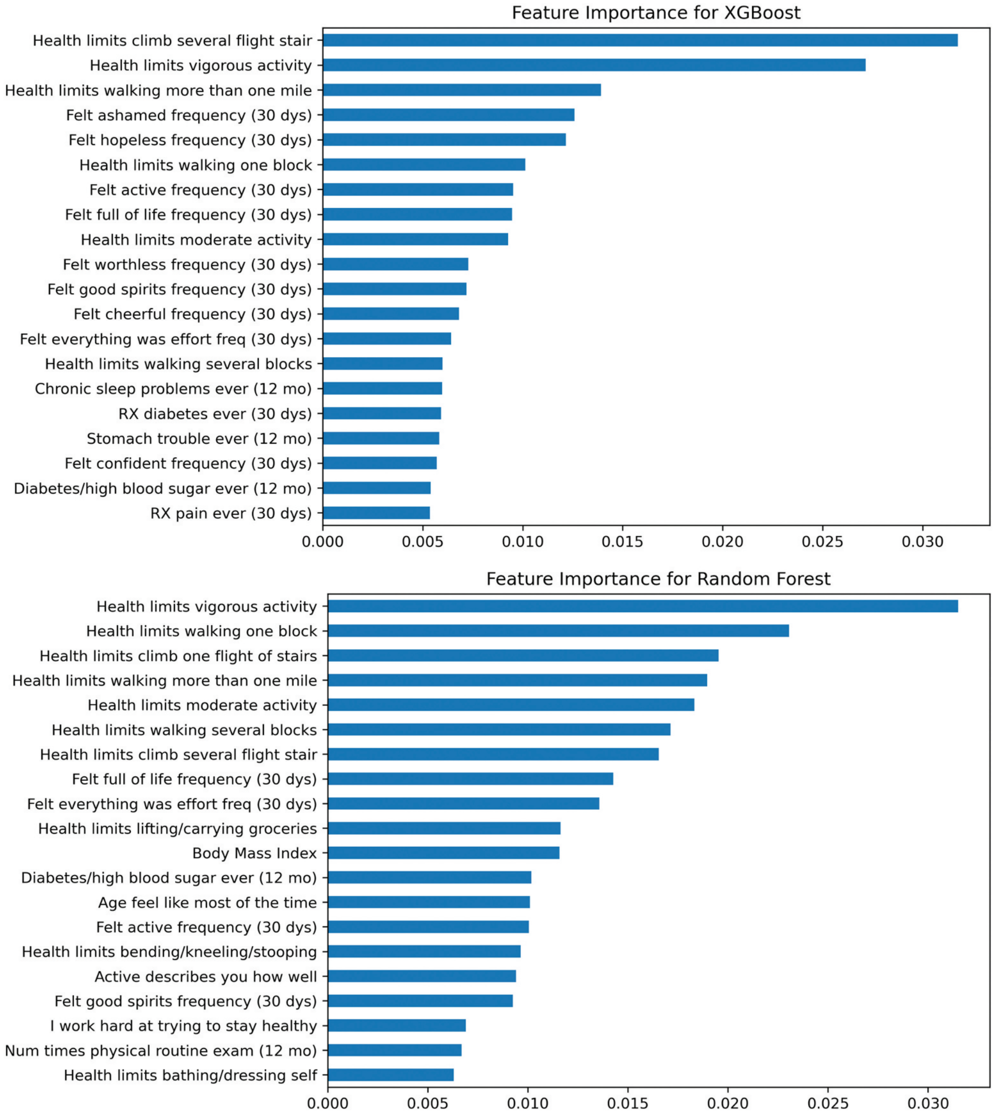

Figure 2. Feature importance for random forest and XGBoost models.

Beyond the common features across both models, important features within the RF model included health-related factors such as body mass index and physical routine exams frequency, and psychological factors such as subjective age (i.e. 'Age feel like most of the time') and personality (e.g. 'active describes you well'). Additional important features within the XGBoost model included taking medications for diabetes or pain, chronic sleep problems, stomach problems, as well as more items on how they felt (e.g. ashamed, hopeless, worthless, cheerful and confident).

Table 2 shows features included in the LASSO regression model. Similar to both RF and XGBoost models, items on functional limitations, and having diabetes or high blood sugar levels, as well as how they felt were included. Unique to the LASSO regression model,

Table 2. Included features in the Final LASSO Regression Model.

Included Features
Health Conditions
Heart trouble suspect/confirmed by doctor
Ever had cancer
Backaches frequency (30 dys)
Sweat frequency (30 dys)
Leaking urine frequency (30 dys)
Intercourse pain/discomfort freq (30dys)
Asthma/bronchitis/emphysema ever (12 mo)
Joint/bone diseases ever (12 mo)
Stomach trouble ever (12 mo)
Constipated all/most ever (12 mo)
High blood press/hypertensn ever (12 mo)
Chronic sleep problems ever (12 mo)
Diabetes/high blood sugar ever (12 mo)
Has chronic pain/persists beyond normal
Medication
Rx heart condition ever (30 days)
Rx hormone therapy ever (30 days)
Rx birth control ever (30 days)
Rx anxiety/depression ever (30 days)
Rx pain ever (30 days)
Taken acetaminophen ever (30 days)
Taken ibuprofen ever (30 days)
Takes multi-vitamins regularly
Takes vitamin C regularly
Takes glucosamine/chondroitin regularly
Functional Limitations
Health limits lifting/carrying groceries
Health limits climb several flights of stair
Health limits climb one flight of stairs
Health limits walking more than one mile
Health limits vigorous activity
Health limits moderate activity
Health Beliefs
Keeping healthy depends on things I do
I work hard at trying to stay healthy
Access to Care
Difficult to get good medical care
Positive/Negative Affect
Felt everything was effort freq (30 dys)
Felt good spirits frequency (30 dys)
Felt satisfied frequency (30 dys)
Felt full of life frequency (30 dys)
Felt active frequency (30 dys)
Psychological Well-Being
Pleased with how life turned out
Demands of everyday life oft get me down
Enjoy make plans for future & make real
Worry about what others think of me
Goal Achievement
No skill/resources reach goal: seek or reconsider goal
Approach to physical health: Stay fit or no worry
Personality
Calm describes you how well
Active describes you how well
Fun learning to walk tightrope
I am a cautious person
Control beliefs
Helpless dealing with problems of life
What happens in life is beyond my ctrl
Able to do things as well as most people
On the whole, I'm satisfied with myself
Like to make plans for future
Can't attain goal, think about oth goals
Keep harmony with others and surroundings
Coping
I concentrate my efforts on doing something about it.
Use Food to Cope
Social Determinants of Health
Own home outright, mortgage, or rent
Highest education
Current work status
Rank standing in community on ladder
Feel safe alone neighborhood at night
Home as nice as most people

taking both prescription and over-the-counter medications were included. Beyond health-related factors, social determinants of health (i.e. education, subjective socioeconomic status, neighborhood safety, and current work status), personality (e.g. 'Calm describes you how well') and control beliefs (e.g. 'I like to make plans for the future'.) were present.

Prediction models without psychological features

Following the high importance of psychological factors in our models, we further tested models that were developed excluding psychological data and compared their

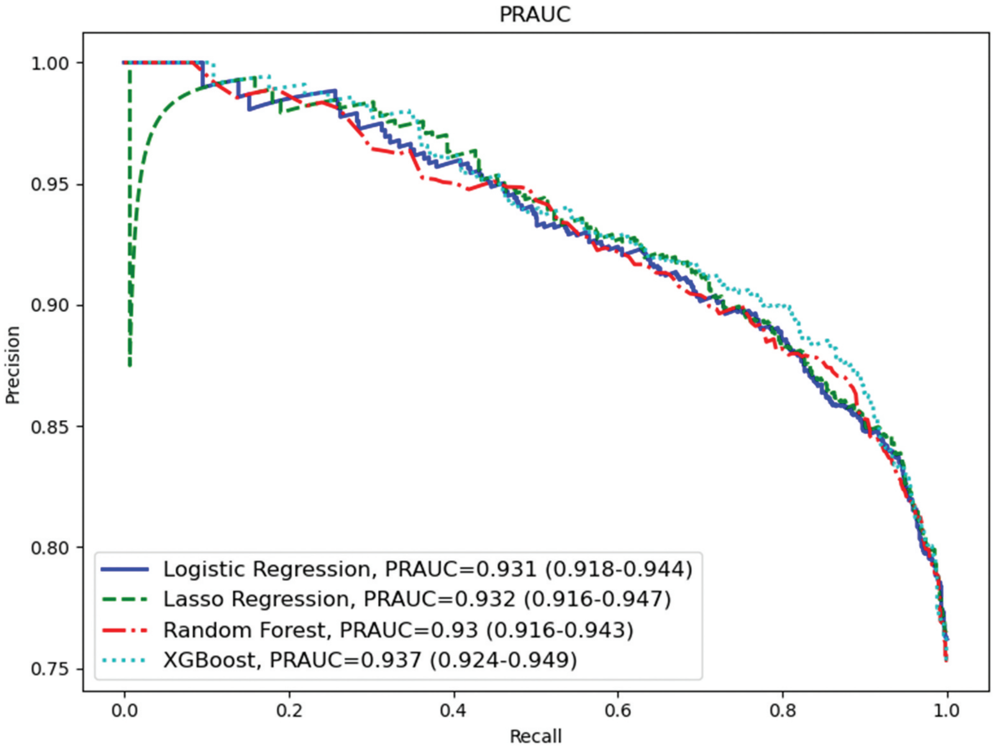

Figure 3. Precision-recall area under curve by prediction Model excluding psychological data.

performance with our main models. Figure 3 displays Area Under Precision–Recall Curves (AUPRCs) of the prediction models developed without psychological data. The models showed high performance (AUPRC > 0.90), yet slightly lower than our main prediction models. XGBoost showed the highest AUPRC (AUPRC = 0.937, confidence interval (CI): [0.924, 0.949]), followed by LASSO (AUPRC = 0.932, CI: [0.916, 0.947]), logistic regression (AUPRC = 0.931, CI: [0.918, 0.944]), and RF (AUPRC = 0.930, CI: [0.916, 0.943]).

Sensitivity analyses

Sensitivity analysis revealed that changing the arbitrary threshold of the outcome did not lead to visible changes on the prediction probability and feature importance of the models (see Supplemental Materials Figure A and B).

Discussion

Using large-scale data encompassing both health-related, sociodemographic, and psychological data, the aim of the study was to develop and compare prediction models via machine learning methods to predict self-rated health. Machine learning models are increasingly used to gain insights into health outcomes, though psychological factors have not been commonly included in such models. We specifically focused on whether prediction models using a wider range of data encompassing both health and psychosocial factors will show strong predictive probabilities, and examined whether psychosocial factors were important features within the models. Supporting our expectations, three of our machine learning algorithms provided great model fit and showed similar important features, both health-related and emotion-related, predicting self-rated health.

Our findings reveal that psychological factors such as emotion, personality, and control beliefs are important features predicting self-rated health. Previous work on prediction models of self-rated health have largely focused on health-related features (Chen et al., 2023; Clark et al., 2021; Gumà-Lao & Arpino, 2023)' overlooking the health impact of psychological factors (Bailis et al., 2003; Choi & Miyamoto, 2022; Cott et al., 1999; Ryff et al., 2015; Stephan et al., 2020; Svedberg et al., 2006). Focusing on health-related factors in developing prediction models may provide a limited perspective on contributors of self-rated health. Findings in fact underscore the complexity and multidimensionality of subjective health, which posit the need to take a multifactorial approach to improving self-rated health.

We also note the consistent importance of emotion-related items across three of our prediction models. This implies that how people feel is strongly linked to how they think of their health as much as their physical functioning. Such finding is not surprising considering the abundant literature on the impact of emotion on health (DeSteno et al., 2013; Ong, 2010) and emotion being an established source of self-rated health (Segerstrom, 2014). Our results further emphasize the importance of emotion in how individuals rate their own health. Furthermore, considering the predictive power of self-rated health in later health outcomes, our study underlines the necessary efforts needed to improve both physical conditions as well as emotional well-being when it comes to improving health overall.

The current study, along with previous studies utilizing prediction models, has implications for how these prediction models can be applied in practical settings. The benefit of machine learning-based prediction models is the potential to identify those who are at risk of low self-rated health. By using prediction models to identify individuals who are likely to report low self-rated health, which in turn could lead to worse long-term health outcomes, we can target such individuals early on to improve their self-reported health. Early identification of those at risk could allow timely intervention focused on both health conditions and emotions, which could in turn improve their subjective health status, and further their long-term health. In addition, our study provides additional insight into how machine learning methods could be utilized in the social sciences. Machine learning methods offer the ability to handle large, complex datasets and identify nuanced patterns among predictors, making them valuable in social sciences for modeling multifaceted constructs like self-rated health. These methods provide an avenue to bridge interdisciplinary approaches, combining health, psychological, and sociodemographic insights. Together, machine learning-based prediction models present new opportunities to reflect the complex nature of self-rated health to both understand and achieve the nuance behind the single-item measure. Current implications will need to be tested and refined through additional research.

This study has several limitations. First, our models were limited to the features available only in MIDUS, specifically focusing on items that are available across the used datasets. Yet, the strength of our study lies in the inclusion of psychological features not available in most population-based datasets. Second, the data only includes middle to older adults, thus the generalizability of the current model to different age ranges need to be tested. Third, our study uses cross-sectional data. Using cross-sectional data precludes the ability to infer causality or track changes in self-rated health over time. Our study captures a single point in time, which limits the capacity to discern whether psychological or health-related factors influence self-rated health, or vice versa. Additionally, cross-sectional designs cannot account for changes in predictors or self-rated health over time. Future research will benefit from longitudinal data to examine whether the predictive probability of the models will be intact in long-term health outcomes. Fourth, we used an arbitrary value for our outcome cutoff to classify high self-rated health and others. Our sensitivity analyses, however, reveal minor changes in feature importance when the outcome threshold was altered to group very good to high health compared to others. Lastly, we did not stratify our sample into subgroups by age, gender, or race.

Using data including a wide spectrum of measures, our prediction models showed high degrees of model fit and identified features of high importance in relation to self-rate health, which encompasses both health and psychological factors. Considering the increasingly active application of machine learning in health research, we thus posit the need to actively consider the role of psychological factors in addition to health-specific factors.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported that there is no funding associated with the work featured in this article.

ORCID

Jeong Ha (Steph) Choi http://orcid.org/0000-0003-4840-9705
Daniel Hong Jung http://orcid.org/0000-0002-2872-0873

References

Awan, S. E., Bennamoun, M., Sohel, F., Sanfilippo, F. M., & Dwivedi, G. (2019). Machine learning-based prediction of heart failure readmission or death: Implications of choosing the right model and the right metrics. *ESC Heart Failure*, *6*(2), 428–435. https://doi.org/10.1002/ehf2.12419

Bailis, D. S., Segall, A., & Chipperfield, J. G. (2003). Two views of self-rated general health status. *Social Science & Medicine*, *56*(2), 203–217. https://doi.org/10.1016/S0277-9536(02)00020-5

Balabaeva, K., & Kovalchuk, S. (2021). Comparison of efficiency, stability and interpretability of feature selection methods for multiclassification task on medical tabular data. In M. Paszynski, D. Kranzlmüller, V. V Krzhizhanovskaya, J. J. Dongarra, & P. M. A. Sloot (Eds.), *Computational science - ICCS 2021* (pp. 623–633). Springer International Publishing.

Benyamini, Y., & Idler, E. L. (1999). Community studies reporting association between self-rated health and mortality: Additional studies & comma; 1995 to 1998. *Research on Aging*, *21*(3), 392–401. https://doi.org/10.1177/0164027599213002

Boscardin, C. K., Gonzales, R., Bradley, K. L., & Raven, M. C. (2015). Predicting cost of care using self-reported health status data. *BMC Health Services Research*, *15*(1). https://doi.org/10.1186/s12913-015-1063-1

Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the precision-recall curve: Point estimates and confidence intervals. In H. Blockeel, K. Kersting, S. Nijssen, & F. Železný (Eds.), *Machine learning and knowledge discovery in databases* (pp. 451–466). Springer.

Brick, T. R., Koffer, R. E., Gerstorf, D., & Ram, N. (2017). Feature selection methods for optimal design of studies for developmental inquiry. *Journals of Gerontology, Series B: Psychological Sciences & Social Sciences*, *73*(1), 113–123. https://doi.org/10.1093/geronb/gbx008

Brim, O. G., Baltes, P. B., Bumpass, L. L., Cleary, P. D., Featherman, D. L., & Hazzard, W. R., Kessler, R. C.,Lachman, M. E.,Markus, H. R.,Marmot, M. G.,Rossi, A. S.,Ryff, C. D. & Shweder, R. A. (2020, September 28). Midlife in the United States (MIDUS 1), 1995-1996. *Inter-University Consortium for Political and Social Research [Distributor]*. https://doi.org/10.3886/ICPSR02760.v19

Chang, W., Liu, Y., Xiao, Y., Yuan, X., Xu, X., Zhang, S., & Zhou, S. (2019). A machine-learning-based prediction method for hypertension outcomes based on medical data. *Diagnostics (Basel, Switzerland)*, *9*(4), 178. https://doi.org/10.3390/diagnostics9040178

Chen, Y., Zhang, X., Grekousis, G., Huang, Y., Hua, F., Pan, Z., & Liu, Y. (2023). Examining the importance of built and natural environment factors in predicting self-rated health in older adults: An extreme gradient boosting (XGBoost) approach. *Journal of Cleaner Production*, *413*, 137432. https://doi.org/10.1016/j.jclepro.2023.137432

Choi, J. H., & Miyamoto, Y. (2022). Cultural differences in self-rated health: The role of influence and adjustment. *The Japanese Psychological Research*, *64*(2), 156–169. https://doi.org/10.1111/jpr.12405

Clark, C. R., Ommerborn, M. J., Moran, K., Brooks, K., Haas, J., Bates, D. W., & Wright, A. (2021). Predicting self-rated health across the life course: Health equity insights from machine learning models. *Journal of General Internal Medicine*, *36*(5), 1181–1188. https://doi.org/10.1007/s11606-020-06438-1

Cornwell, E. Y., & Waite, L. J. (2009). Social disconnectedness, perceived isolation, and health among older adults. *Journal of Health and Social Behavor*, *50*(1), 31–48. https://doi.org/10.1177/002214650905000103

Cott, C. A., Gignac, M. A. M., & Badley, E. M. (1999). Determinants of self rated health for Canadians with chronic disease and disability. *Journal of Epidemiology & Community Health*, *53*(11), 731–736. https://doi.org/10.1136/jech.53.11.731

Darabi, N., Hosseinichimeh, N., Noto, A., Zand, R., & Abedi, V. (2021). Machine learning-enabled 30-day readmission Model for stroke patients. *Frontiers in Neurology*, *12*, 638267. https://doi.org/10.3389/fneur.2021.638267

Desalvo, K. B., Jones, T. M., Peabody, J., Mcdonald, J., Fihn, S., Fan, V., He, J., & Muntner, P. (2009). *Health care expenditure prediction with a single item, self-rated health measure*. http://journals.lww.com/lww-medicalcare

DeSteno, D., Gross, J. J., & Kubzansky, L. (2013). Affective science and health: The importance of emotion and emotion regulation. *Health Psychology*, *32*(5), 474–486. https://doi.org/10.1037/a0030259

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00516-9

Fayers, P. M., & Sprangers, M. A. G. (2002). Understanding self-rated health. *Lancet*, *359*(9302), 187–188. https://doi.org/10.1016/S0140-6736(02)07466-4

Franks, P., Gold, M. R., & Fiscella, K. (2003). Sociodemographics, self-rated health, and mortality in the US. *Social Science & Medicine*, *56*(12), 2505–2514. https://doi.org/10.1016/S0277-9536(02)00281-2

Göbl, C. S., Bozkurt, L., Tura, A., Pacini, G., Kautzky-Willer, A., Mittlböck, M., & Friede, T. (2015). Application of penalized regression techniques in modelling insulin sensitivity by correlated metabolic parameters. *PLOS ONE*, *10*(11), e0141524–e0141524. https://doi.org/10.1371/journal.pone.0141524

Gumà-Lao, J., & Arpino, B. (2023). A machine learning approach to determine the influence of specific health conditions on self-rated health across education groups. *BMC Public Health*, *23*(1). https://doi.org/10.1186/s12889-023-15053-8

Idema, C. L., Roth, S. E., & Upchurch, D. M. (2020). Weight perception and perceived attractiveness associated with self-rated health in young adults. *Preventive Medicine*, *120*(July 2018), 34–41. https://doi.org/10.1016/j.ypmed.2019.01.001

Idler, E. L., Russell, L. B., & Davis, D. (2000). Survival, functional limitations, and self-rated health in the NHANES I epidemiologic follow-up study, 1992. *American Journal of Epidemiology*, *152*(9), 874–883. https://doi.org/10.1093/aje/152.9.874

Jung, D., Pollack, H. A., & Konetzka, R. T. (2023). Predicting hospitalization among medicaid home- and community-based services users using machine learning methods. *Journal of Applied Gerontology*, *42*(2), 241–251. https://doi.org/10.1177/07334648221129548

Jylhä, M., Guralnik, J. M., Ferrucci, L., Jokela, J., & Heikkinen, E. (1998). Is self-rated health comparable across cultures and genders? *Journals of Gerontology - Series B Psychological Sciences & Social Sciences*, *53*(3), 144–152. https://doi.org/10.1093/geronb/53B.3.S144

Krause, N. M., & Jay, G. M. (1994). What do global self-rated health items measure? *Medical Care*, *32*(9), 930–942. https://doi.org/10.1097/00005650-199409000-00004

Latham, K., & Peek, C. W. (2013). Self-rated health and morbidity onset among late midlife U.S. Adults. *Journals of Gerontology, Series B: Psychological Sciences & Social Sciences*, *68*(1), 107–116. https://doi.org/10.1093/geronb/gbs104

Liu, W., Stansbury, C., Singh, K., Ryan, A. M., Sukul, D., Mahmoudi, E., Waljee, A., Zhu, J., Nallamothu, B. K., & Liu, N. (2020). Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. *PLOS ONE*, *15*(4), e0221606–e0221606. https://doi.org/10.1371/journal.pone.0221606

Molarius, A., & Janson, S. (2002). Self-rated health, chronic diseases, and symptoms among middle-aged and elderly men and women. *Journal of Clinical Epidemiology*, *55*(4), 364–370. https://doi.org/10.1016/S0895-4356(01)00491-7

OECD. (2023). *Health at a glance, 2023: OECD indicators*. OECD Publishing. https://doi.org/10.1787/7a7afb35-en

Ong, A. D. (2010). Pathways linking positive emotion and health in later life. *Current Directions in Psychological Science*, *19*(6), 358–362. https://doi.org/10.1177/0963721410388805

Petrazzini, B. O., Naya, H., Lopez-Bello, F., Vazquez, G., & Spangenberg, L. (2021). Evaluation of different approaches for missing data imputation on features associated to genomic data. *BioData Mining*, *14*(1). https://doi.org/10.1186/s13040-021-00274-7

Radler, B. T. (2014). The midlife in the United States (MIDUS) series: A national longitudinal study of health and well-being. *Open Health Data*, *2*(1). https://doi.org/10.5334/jophd.ai

Ryff, C. D., Almeida, D. M., Ayanian, J. Z., Carr, D. S., Cleary, P. D., & Coe, C., Davidson, R. J., Krueger, R. F,Lachman, M. E.,Marks, N. F.,Mroczek, D. K. & Williams, D. R. (2021, September 15). Midlife in the United States (MIDUS 2), 2004-2006. *Inter-University Consortium for Political and Social Research [Distributor]*. https://doi.org/10.3886/ICPSR04652.v8

Ryff, C. D., Radler, B. T., & Friedman, E. M. (2015). Persistent psychological well-being predicts improved self-rated health over 9-10 years: Longitudinal evidence from MIDUS. *Health Psychology Open*, *2*(2). https://doi.org/10.1177/2055102915601582

Saito, T., Rehmsmeier, M., & Brock, G. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/JOURNAL.PONE.0118432

Segerstrom, S. C. (2014). Affect and self-rated health: A dynamic approach with older adults. *Health Psychology*, *33*(7), 720–728. https://doi.org/10.1037/a0033506

Sofaer, H. R., Hoeting, J. A., Jarnevich, C. S., & McPherson, J. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577. https://doi.org/10.1111/2041-210X.13140

Stephan, Y., Sutin, A. R., Luchetti, M., Hognon, L., Canada, B., & Terracciano, A. (2020). Personality and self-rated health across eight cohort studies. *Social Science & Medicine*, *263*, 263. https://doi.org/10.1016/j.socscimed.2020.113245

Svedberg, P., Bardage, C., Sandin, S., & Pedersen, N. L. (2006). A prospective study of health, life-style and psychosocial predictors of self-rated health. *European Journal of Epidemiology*, *21* (10), 767–776. https://doi.org/10.1007/s10654-006-9064-3

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2018). *Introduction to data mining* (2nd Edition), 2nd ed.). Pearson.

Taninaga, J., Nishiyama, Y., Fujibayashi, K., Gunji, T., Sasabe, N., Iijima, K., & Naito, T. (2019). Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study. *Scientific Reports* *9*(1), 1–9. https://doi.org/10.1038/s41598-019-48769-y

Watson, D., & Pennebaker, J. W. (1989). Health complaints, stress, and distress: Exploring the central role of negative affectivity. *Psychological Review*, *96*(2), 234–254. https://doi.org/10.1037/0033-295X.96.2.234

Wu, S., Wang, R., Zhao, Y., Ma, X., Wu, M., Yan, X., & He, J. (2013). The relationship between self-rated health and objective health status: A population-based study. *BMC Public Health*, *13* (1). https://doi.org/10.1186/1471-2458-13-320

Wuorela, M., Lavonius, S., Salminen, M., Vahlberg, T., Viitanen, M., & Viikari, L. (2020). Self-rated health and objective health status as predictors of all-cause mortality among older people: A prospective study with a 5-, 10-, and 27-year follow-up. *BMC Geriatrics*, *20*(1). https://doi.org/10.1186/s12877-020-01516-9