Economics Bulletin

Volume 46, Issue 1

Worker and workplace predictors of self-reported health status: An application of econometrics and machine learning

M. Ryan Haley
University of Wisconsin Oshkosh

Abstract

This paper applies an array of econometric and machine learning techniques to individual-level data from the 2008 National Study of the Changing Workforce to identify worker and workplace features that appear to be strong predictors of self-reported health status. Marginal effects and log-odds ratios are discussed, as is prediction performance. Of all the methods used, the (gradient) boosted-tree and support-vector-machine models delivered the best prediction accuracy. In assessing variable importance, several features consistently stood out cross the array of analyses: depression, sleep difficulties, home stress, race, Hispanic ethnicity, life satisfaction, work-family conflict, education, and access to specific types of flexible working arrangements. Earnings, age, and female variables, while not significant in the baseline analyses, emerged as relevant in several sensitivity analyses. Some of these findings are policy items (e.g., flexible working arrangements, work-family conflict issues as well as race, gender, and ethnicity issues), which might be considered by firms, unions, and/or policy makers to improve self-reported health results moving forward; other significant features such a depression and sleep difficulties are more within the individual's purview to remediate.

Citation: M. Ryan Haley, (2026) "Worker and workplace predictors of self-reported health status: An application of econometrics and machine learning", *Economics Bulletin*, Volume 46, Issue 1, pages 61-70
Contact: M. Ryan Haley - haley@uwosh.edu.
Submitted: March 20, 2025. **Published:** March 30, 2026.

1. Introduction and Background

Self-reported health is an important indicator that frequently appears in survey research. It often correlates with actual health status (Wu et al., 2013; Balaj, 2022) and can even predict future health outcomes, including morbidity and mortality in certain settings. Therefore, ascertaining which worker and workplace features might meaningfully impact self-reported health is a worthy research endeavor, and the one pursued herein. Understanding which factors significantly impact self-reported health can assist firms, unions, policymakers, and individuals in promoting positive health practices and behaviors while reducing negative ones. The goal is to complete this inquiry by using traditional statistical/econometric methods alongside more recently developed machine learning techniques. This combination of diverse methods offers an especially rich estimation strategy that is often absent in prior studies.

Self-reported health has been widely studied in many areas of academic inquiry, among them Economics, Management, Medicine, Psychology, and Public Health. For example, Johansson et al. (2020) studied how self-reported health compared to biomarkers, and how each interacted with unemployment. In another labor-related study, Gevaert et al. (2021) studied how self-reported health intermixed with what they call "employment quality"; they found strong associations between mental well-being and self-reported general health. Gender and self-reported health have also been studied at some length. For example, Roxo et al. (2021) found gender inequality in self-reported health across 27 European countries, wherein females typically self-reported lower levels of health; similar inquiries include Hosseinpoor et al. (2012) and Harnois and Bastos (2018).

Regarding estimation strategies: several leading economists have begun advocating the merits of infusing machine learning techniques into econometric settings; see, for example, Varian (2014), Athey (2018), or Athey and Imbens (2019). Kino et al. (2021) suggest value in exploring health issues using machine learning techniques. Qin et al. (2020) studied health status in older populations and found that artificial neural networks (ANNs) gave the best prediction accuracy; they also noted that machine learning was especially useful for capturing non-linear predictor associations with health outcomes. Like Kino et al. (2021), they likewise suggested that machine learning may offer more than conventional methods in health inquiries. In research similar to the present study, Clark et al. (2021) applied several machine learning methods to the Behavioral Risk Factor Surveillance System **(**BRFSS) data in search of features that drive self-reported health; several of their key findings are corroborated herein (see Estimation Section for more details). Olsson et al. (2022) performed a similar exercise among older men in Sweden and found that sleep and depression lowered self-reported health. Gumà-Lao and Arpino (2023) did a similar study with older people bracketed by educational attainment level; they likewise found depression lowers self-reported health. Wallace et al. (2019) used machine learning methods to ascertain the impact of sleep issues on health status, though they used more medically-focused health outcomes like cardiovascular mortality.

Flexible working arrangements (FWAs) have also been studied at length and some research suggests they may interplay with self-reported health; e.g., Pollmann-Schult (2018) and Kim et al. (2020). The potential advantages to FWAs are many; e.g., Bloom and van Reenen (2006) found that FWAs may increase productivity; Baughman et al. (2003) found a reduction in turnover in the

presence of flexible sick leave and childcare assistance policies; and Cotti et al. (2014) found FWAs helped increase job satisfaction and that they interplayed with workplace stress, union membership, and job control. Finally, health issues such as stress, depression, and sleep troubles have been studied alongside life satisfaction, job satisfaction, and FWAs; e.g., Halpern (2005), Blackmore et al. (2007), Grzywacz et al. (2007), Haley and Miller (2015, 2023), Cotti et al. (2017), Lee at al. (2020), and Kudrnáčová and Kudrnáč (2023).

Taken together, this methodologically diverse background literature motivates the research question and research methods used in the present paper, and also helps outline a proper set of covariates/features to consider in such an analysis. Data sourcing and covariate/feature selection is the focus of the next section.[1]

2. Data and Feature Selection

The 2008 NSCW survey provides a cross-sectional representative sample of the United States (US) workforce and includes detailed information about employees and employment conditions. Individual workers are the units of observation. Response fields vary by question type, some requiring a fill-in response (e.g., age), while others use indicator, Likert, or Likert-like scales. Some questions include responses such as 'don't know', 'refuse', or 'not applicable', which were treated as missing. The focus is on respondents that work more than 20 hours per week and who are not self-employed, which left 2152 observations. In several circumstances, Likert or Likert-like variables were converted into indicator variables.[2]

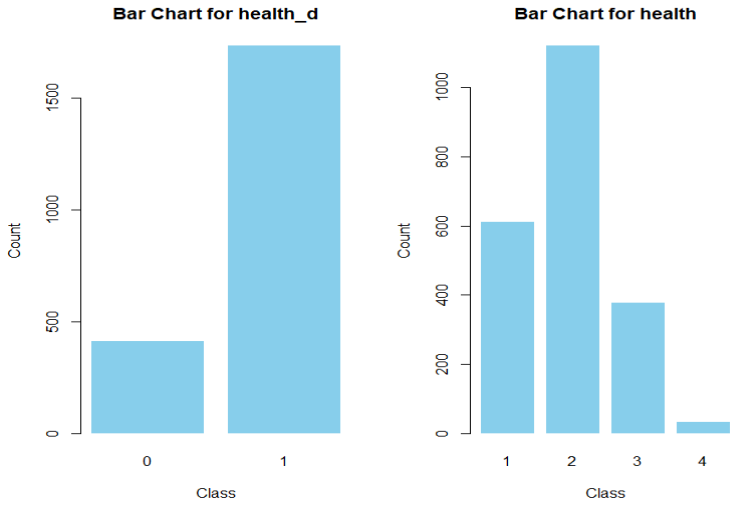

Figure 1: This figure reports the responses for the binary (**health_d**) and Likert-scale (**health**) versions of the dependent variable, which gauges self-reported health status. The proportion for the binary version is 0.807 and the mean of the Likert version is 1.925. The binary version = 1 if reported health is good or excellent; 0 otherwise. Sample size = 2152.

[1] The term "covariate" is likely familiar to economic readers, while machine learning readers will likely be more familiar with the term "feature."
[2] Full details about the data set appear in the Data Appendix.

The NSCW's primary self-reported health status question is as follows:

> QPW16C: How would you rate your current state of health – excellent (1), good (2), fair (3), or poor (4)?

The responses to this question are summarized in two ways; first, in its original four-class form and second, in a binary format; see Figure 1.

Table 1: Estimation Results*

Feature	OLS Estimate	t value	Logit Estimate	z value	LASSO	Ridge
intercept	0.907	5.677	2.452	2.194	1.601	1.897
flex1	0.014	0.71	0.094	0.725	0.08	0.147
flex2	0.093	3.86	0.553	3.663	0.295	0.309
flex3	-0.004	-0.196	-0.02	-0.147	0	-0.031
flex4	-0.005	-0.201	-0.037	-0.205	0	0.039
flex5	-0.024	-1.352	-0.194	-1.568	0	-0.075
age	-0.001	-1.117	-0.006	-1.115	-0.001	-0.004
female	0.02	1.154	0.133	1.093	0	0.034
hispanic	-0.103	-2.711	-0.621	-2.642	-0.318	-0.412
physical_job	0.02	1.117	0.159	1.259	0.073	0.126
vacation	-0.005	-0.24	-0.035	-0.228	0	0.106
union	-0.017	-0.755	-0.111	-0.724	0	-0.072
job_sat	-0.007	-0.219	-0.083	-0.423	0	0.084
ln_earn	-0.021	-0.512	-0.177	-0.623	0	-0.12
fast_job	0.034	1.747	0.228	1.731	0.186	0.212
depress1	-0.085	-3.818	-0.463	-3.436	-0.387	-0.403
kids	0.01	0.475	0.073	0.517	0.078	0.123
eldercare	-0.005	-0.312	-0.037	-0.317	0	0.011
race_black	-0.13	-4.375	-0.741	-4.157	-0.605	-0.607
married_part	0.004	0.201	0.03	0.229	0	0.005
wfc_index_d	-0.098	-2.951	-0.498	-2.53	-0.45	-0.472
life_sat_d	0.079	4.227	0.644	4.635	0.67	0.576
high_school	-0.069	-3.457	-0.526	-3.581	-0.437	-0.421
public_emp	0.011	0.501	0.093	0.604	0	0.063
night_shift	-0.003	-0.067	-0.012	-0.041	0	-0.006
home_stress_d	-0.057	-2.121	-0.294	-1.837	-0.257	-0.321
sleep_index_d	-0.131	-6.643	-0.791	-6.297	-0.697	-0.618

*This table contains OLS marginal effects (and t-values), logit log-odds estimates (and z values), and estimates from LASSO and Ridge regularized regressions. Accuracy gathered from the confusion matrices was 0.82 for all three models (excluding OLS). Shading indicates significance at the 5% level (or less) for OLS and logit. Sample size = 2152.

Selecting covariates and controls is always a delicate matter. To navigate these decisions, theoretical relevance of variables used in prior research was weighed heavily, and careful thought was given to the probable direction of "causality" between candidate covariates and the dependent

variable. In addition, the usual correlation and variance inflation factor (VIF) assessments were applied to the pool of candidate covariates. These approaches hopefully minimized the probability of any meaningful econometric shortcomings in the specifications. After implementing these screens, numerous covariates remained, which spanned workers' physical characteristics, anxiety, depression, life satisfaction, race, ethnicity, marital status, education, sleep quality, industry, occupation, and region; see the Data Appendix for full details. These controls are frequently used in labor/health-economic studies of this type.

3. Estimation and Learning

The baseline estimation strategy is a standard logit specification using **health_d** from Figure 1, the dichotomized version of the NSCW's health status question, as the response variable. The primary estimation strategy is a binary logit model:

$$P(\text{health}_d=1) = \frac{1}{1+\exp[-(\beta_0+\sum_{k=1}^{K}\beta_k X_k)]}$$

where $\text{health}_d = 1$ if self-reported health is "good" or "excellent," and X_k represents the covariates listed in Table 1. All covariates were included simultaneously in the baseline specification. Regularized regressions (LASSO and Ridge) were applied to the same set of covariates to assess robustness.[3] The baseline results appear in Table 1. The regularized results largely agreed with the OLS and logit results. Below is a specific discussion of each of the primary findings (with signs) from Table 1.

- **flex2 (+):** This suggests that the presence of short-notice schedule flexibility (flex2) was associated with higher health status; this is consistent with Joyce et al. (2010) and Erhel et al. (2024). That the other FWAs were not significant is unsurprising based on Haley and Miller (2015), which reported that the flexibility of an FWA (i.e., the flexibility of the flexibility) often matters most to health-oriented outcomes.

- **hispanic (-):** This suggests that Hispanic ethnicity is correlated with lower levels of self-reported health; this is consistent with Gandhi et al. (2020) and Mahajan et al. (2021).

- **depress1 (-):** This suggests that depression reduces self-reported health; this is consistent with many prior studies such as Ishida et al. (2020), Vaingankar et al. (2020), Clark et al. (2021), Olsson et al. (2022), and Gumà-Lao and Arpino (2023).

[3] Regularized regression techniques extend traditional regression by adding a penalty term to the size of coefficients. LASSO uses an absolute-value penalty, which can shrink some coefficients to zero, effectively performing variable selection. Ridge uses a squared penalty, which shrinks coefficients toward zero without eliminating variables, improving stability when predictors are highly correlated. Both methods help prevent overfitting and improve predictive performance in models with many covariates.

- **race_black (-):** This suggests that black respondents report lower levels of health; this is consistent with many prior findings, among then Whiting and Bartle-Haring (2022), Gandhi et al. (2020), McNeil et al. (2020), and Mahajan et al. (2021).

- **wfc_index_d (-):** This result suggests that respondents with higher levels of work-family conflict tend to report lower levels of health; this is consistent with Mensah and Adjei (2020).

- **lif_sat_d (+):** This suggests that respondents with higher levels of life satisfaction tend to report better health; this is consistent with Koivumaa-Honkanenet et al. (2000), Kööts–Ausmees and Realo (2015), and Kim et al. (2021).

- **high_school (-):** This suggests that respondents with lower levels of education tend to report lower health values; this is consistent with Subramanian et al. (2010), Molina (2016), and Clark et al. (2021), among many others.

- **home_stress_d (-):** This suggests that higher levels of stress lower self-reported health; this is a consistent with Dunn et al. (2021).

- **sleep_index_d (-):** This suggests that higher levels of sleep difficulty lower self-reported health levels; this is consistent with Haley and Miller (2015), Olsson et al. (2022), and Kudrnáčová and Kudrnáč (2023).

While many of the results in Table 1 conform with the background literature, there were several variables that, somewhat surprisingly, did not emerge as significant. First, absolute log earnings did not play a role in health status in the baseline analyses, though Boyce (2010) and Boodoo et al. (2014) have argued, in similar settings, that relative income matters more than absolute income. Second, some research reports that job satisfaction is positively correlated with self-reported health (e.g., Fischer and Sousa-Poza, 2009**),** though other research suggests this association is less clear (e.g., Svensen et al., 2007); regardless, job satisfaction was not significant in the baseline results. Third, age is sometimes reported to affect health, and sometimes not (e.g., Lin et al. 2022); age did not present as significant in the baseline analyses. Fourth, female sex is sometimes reported to correlate with lower self-reported health (e.g., Clark et al., 2021); such was not the case in the baseline results. Fifth, marital status is often associated with higher self-reported health levels (e.g., Lindström, 2009; Zella, 2017); this association was not evident in the 2008 NSCW data. These mixed findings become a focal point in the sensitivity analyses to follow.

4. Sensitivity Analyses

It is of course prudent to assess the fragility of the results in Table 1. To do so, an array of alternative specifications and estimation strategies were explored, which are consolidated below into three sensitivity analyses.

Sensitivity Analysis #1: The Likert-scaled version of the dependent variable (see Figure 1) was applied using ordered logit. The results, which appear in Figure 2, are similar to the baseline

findings in Table 1, but also pick up the potential importance of age, union membership, and a second FWA (**flex1**).

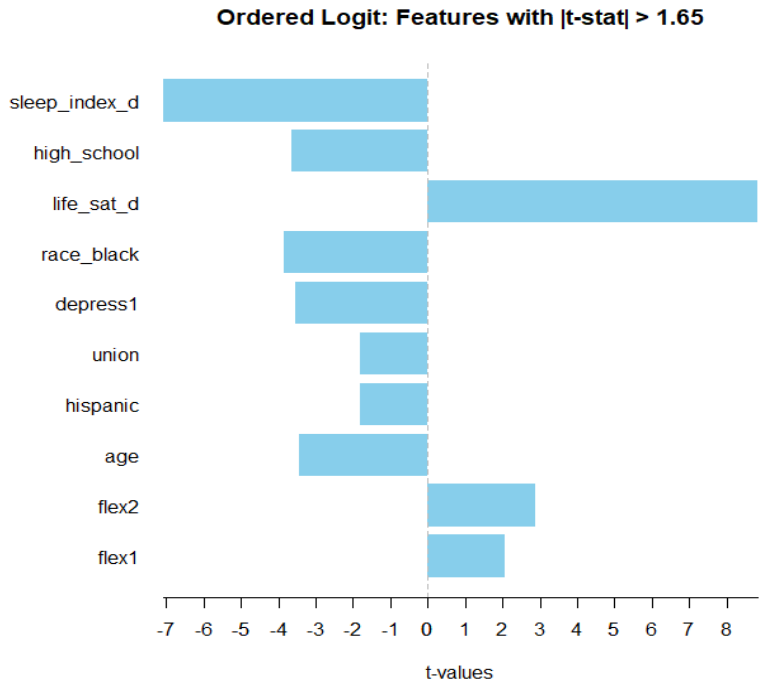

Figure 2: This chart reports the statistically significant features (at the 10% level) from the ordered-logit estimation, which used the four-class version of the health variable reported in Figure 1. Sample size = 2152.

Sensitivity Analysis #2: Standard OLS was applied to the dummy version of the dependent variable from Figure 1 using the R language's leaps library, which performs an exhaustive specification search using the covariates provided and summarizes the results with a BIC chart. The results appear in Figure 3. The charts display which covariates appear most often in the best models found in repeated applications of OLS across different subsets of covariates. The results largely parallel the baseline results in Table 1.

Sensitivity Analysis #3: Four additional machine learning algorithms were applied to the NSCW data, from which variable importance scores are reported for the 10 most important features; see Figure 4. The four methods explored herein were as follows:

- **Random Forests:** This technique builds many decision trees on random subsets of the data and predictors. Each tree makes a prediction, and the forest aggregates these predictions (e.g., majority vote for classification). This approach reduces overfitting and handles complex interactions well, making it useful for real-world data.

- **Gradient Boosted Trees:** This method constructs trees sequentially, where each new tree focuses on correcting errors from previous ones. This iterative process often leads to very high predictive accuracy, especially when relationships between variables are non-linear or involve subtle interactions.

- **Support Vector Machines (SVM):** This method finds the best boundary (hyperplane) that separates observations into classes. SVMs work well in high-dimensional spaces and can use "kernels" to capture non-linear patterns without explicitly transforming the data.

- **Naïve Bayes:** This is a simple probabilistic model based on Bayes' theorem. It assumes predictors are conditionally independent given the outcome—a strong assumption, but one that often works surprisingly well in practice. Naïve Bayes is fast, interpretable, and effective for high-dimensional data.

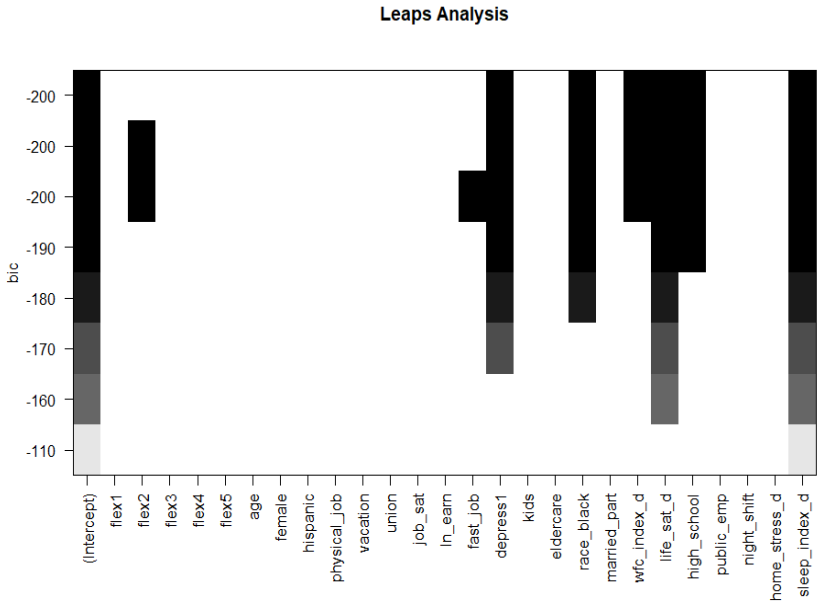

Figure 3: This chart shows the results of the R language's leaps library. The covariates that emerged as important here compare favorably to those identified by the baseline OLS and logit models. Sample size = 2152.

Interestingly, these machine learning analyses pick up many of the mixed results that were discussed in the prior section. For example, age, log earnings, and female sex were ranked highly by several of the machine learning models; the Naïve Bayes model picked up the importance of marital status; and the SVM picked up the importance of job satisfaction. The baseline results plus the sensitivity results more comprehensively outline, one might argue, the worker and workplace features germane to self-reported health status. These results also suggest that using a narrow spectrum of traditional econometric methods may lead to empirical conclusions that are not as robust as researchers or policy makers might desire.

5. Summary, Discussion, and Conclusions

This paper studies how various worker and workplace characteristics predict self-reported health using data from the 2008 NSCW. Some of these results are actionable by firms via corporate

wellness programs and workplace-positive management practices (such as offering attractive FWAs and helping to minimize work-family conflict), any or all of which may improve self-reported health. Other covariates, such as depression, sleep troubles, and home stress are more within the purview of individuals to adjust; but they can take some solace in knowing that ameliorating these challenges will likely lead to better health outcomes. Moreover, better self-reported health correlates with better actual health, which underscores the importance of pursuing these firm- and individual-level pro-health actions.

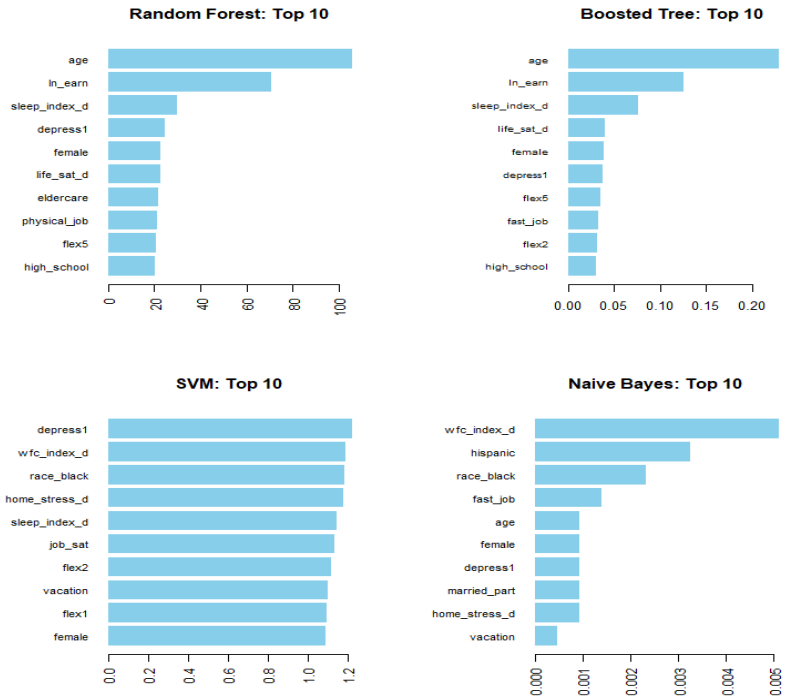

Figure 4: This chart reports the results of variable importance analyses from random forest, (gradient) boosted tree, SVM, and Naïve Bayes models. The 10 most important features are reported for each method. The prediction accuracy rates from the confusion matrices were 0.81, 0.98, 0.89, 0.73, respectively. Sample size = 2152.

A theme of this paper was to use a blend of econometric and machine learning techniques to create a more diversified estimation environment. In this setting, it appears to have added value, as the baseline regression-based models had some intuitive weaknesses (i.e., missed some features that probably do impact health reporting) that the sensitivity analyses picked out, which in turn helped create a richer understanding of which worker and workplace characteristics appear to drive self-reported health. More specifically, covariates such as age and earnings emerged as important in machine learning models but were statistically insignificant in the baseline regression analyses. This divergence is not entirely unexpected and likely stems from fundamental differences in the objectives and mechanics of these disparate methods. Traditional regression models estimate marginal effects under linearity and additivity assumptions, focusing on statistical significance for inference. In contrast, machine learning algorithms prioritize predictive accuracy and can capture complex non-linear relationships and high-order interactions among covariates. Consequently, a variable may exhibit weak marginal effects in a linear model yet contribute substantially to prediction when combined with other features in a non-linear framework. This distinction underscores the complementary nature of econometric and machine learning approaches in health

research (Athey and Imbens, 2019; Varian, 2014). Also of note is how machine learning methods like random forests, gradient boosted trees, SVMs, and artificial neural networks (ANNs) capture non-linear aspects in a much more general and flexible way than a parametric method like logit. As such, one conclusion of this research is that machine learning methods serve an important purpose in these types of health studies, and could (should?) be fruitfully applied more broadly moving forward.

Another noteworthy result concerns the Hispanic paradox, which refers to the observation that Hispanic Americans often experience lower mortality and longer life expectancy than non-Hispanic Whites, despite socioeconomic disadvantages. However, this advantage applies primarily to objective outcomes such as mortality. In contrast, Hispanics frequently report lower self-rated health compared to Whites, a pattern documented in national surveys and attributed to cultural differences in health perception and reporting norms. Accordingly, the finding that Hispanic ethnicity is associated with lower self-reported health (see Table 1 and Figure 2) does not necessarily contradict the paradox; rather, it most likely reflects the distinction between subjective health assessments and actual mortality outcomes (see, for example, Gandhi et al., 2020; Mahajan et al., 2021).

Several topics for future research are apparent. First, further explorations into absolute vs. relative wages would be of interest. Second, it appears that some FWAs are more able to impact health status than others; thus, studying FWA types more surgically would be of interest. Third, Boerma et al, (2016) and Molina (2016) report that self-report health varies by gender; some evidence for this was found in the current paper as well. Similarly, there is evidence that race and ethnicity are meaningful drivers of self-reported health. As such, detailed gender-specific, race-specific, and ethnicity-specific inquiries would be good follow ups to the present paper.

Data Appendix

This appendix contains details about the variables used in the paper.

- **flex_import:** =1-4.
    - QWC48G: Imagine that you were looking for a new job. How important would [each of] the following [things] be in deciding to take that job: Having the flexibility I need to manage my work and personal or family life ---
        - 1: extremely important
        - 2: very important
        - 3: somewhat important
        - 4: not important.
- **flex_import_d** = 1 if flex_import <=2; 0 otherwise.
- **female:** =1 if respondent is female; 0 if not.
    - QSC8 (1=Male; 2=Female)
- **age:** Age of respondent in years.
    - QPD1: First, may I ask how old you are? (range = 18-100)
- **Health_d:** =1 if reported health is good or excellent; 0 otherwise.
    - QPW16C: How would you rate your current state of health – excellent (1), good (2), fair (3), or poor (4)?
- **sleep_index:** Is an index constructed by averaging three question on sleep issues, all of which are measured on the same five point Likert scale (1: never - 5: very often).
    QPW2: How often have you had trouble sleeping to the point that it affected your performance on and off the job?
    - QPW7A: How often have you had trouble falling asleep when you go to bed?
    - QPW7B: How often have you awakened before you wanted to and had trouble falling back asleep?
- **sleep_index_d**: = 1 for response to sleep_index >=3; 0 otherwise.
- **kids** (kidles13): =1 if any child < 13 in household for more than 1/2 the year; 0 otherwise
- **eldercare:** =1 if respondent has provided eldercare with past five years; 0 otherwise
    - QEC7: Within the past 5 years have you provided special attention or care for a relative or in-law 65 years old or older -- helping with things that were difficult or impossible for them to do themselves? (1:yes, 2:no)
- **race:** QPD4: What is your race? Dummy variable for each of the following:
        - 1: White
        - 2: Black
        - 3: Native American or Alaskan Native
        - 4: Asian, Pacific Islander, or Indian (from India)
        - 5: Other, including mixed
- **race_white:** = 1 if white; 0 otherwise.
- **race_black:** = 1 if black; 0 otherwise.
- **hispanic:** =1 if respondent is Hispanic; 0 if not.
    - QPD3: Do you identify yourself as Hispanic? (1:yes, 2:no)
- **marital_status:**

    - QEN8: Are you presently married for the first time, remarried following a divorce, living with someone as a couple, single and never married, divorced, widowed or separated?
        - 1: Married for the first time
        - 2: Remarried
        - 3: Living with someone as a couple
        - 4: Single and never married and not living with someone as a couple
        - 5: Divorced and not living with someone as a couple
        - 6: Widowed and not living with someone as a couple
        - 7: Separated and not living with someone as a couple
- **married_part**: =1 if marital_status <=3; 0 otherwise.
- **education** (QPD2): What is the highest level of schooling you have completed? A dummy for each of the following:
    - 1: Less than high school
    - 2: High school or GED
    - 3: Trade or technical school beyond high school
    - 4: Some college
    - 5: Two-year Associate's Degree
    - 6: Four/five-year Bachelor's Degree
    - 7: Some college after BA or BS but without degree
    - 8: Professional degree in medicine, law, dentistry
    - 9: Master's Degree or Doctorate
- **high_school:** =1 if education <=5; 0 otherwise.
- **life_sat_d:** =1 if very satisfied; 0 otherwise.
    - QPW10: All things considered, how do you feel about your life these days? (1: very satisfied- 4: very dissatisfied).
- **stress_index:** =1if stress index is greater than or equal to three. The stress index is constructed by taking the average response to the following four questions all of which are measured on the same five point Likert scale (1: never - 5: very often):
    - QPW3: How often have you felt nervous and stressed?
    - QPW4: How often have you felt that you were unable to control the important things in your life?
    - QPW6: How often have you felt that things were going your way?
    - QPW7: How often have you felt that difficulties were piling up so high that you could not overcome them?
- **stress_index_d**: =1 if response to stress_index >= 3; 0 otherwise.
- **home_stress:** =1 if extremely (1), very (2), or somewhat (3) stressful.
    - QPW17: Not thinking about work, how stressful has your personal and family life been in recent months? (1:extremely stressful – 5:not stressful at all)
- **depress1:** =1 if yes; 0 if not
    - QPW8: During the past month, have you been bothered by feeling down, depressed, or hopeless? (1:yes, 2:no)
- **earnings:** = ln[min(earnings) + earnings], to ensure a real number.
    - QSS1: How much did you personally earn in all of LAST YEAR , including bonuses, from all paid employment before taxes? (range = 0 – infty)

- **fast_job** (QWC2): My job requires that I work very fast.
  - (1:strongly agree – 4:strongly disagree)
- **physical_job** (QWC13A): My job requires a lot of physical effort.
  - (1:strongly agree – 4:strongly disagree)
- **perf_pay**: =1 if yes; 0 otherwise.
  - QSS3: At your main job are pay increases, bonuses, and promotions directly and clearly related to your performance? (1:Yes, 2:No)
- **job_sat** (QWC38): All in all, how satisfied are you with your job? (1: very satisfied - 4: not at all satisfied)
- **vacation:** =1 if yes; 0 otherwise.
  - QBP14: Do you receive any PAID vacation days at your main job? (1:yes, 2:no)
- **firm_size** (QEB42B): Approximately how many people are employed by your company or organization at YOUR work location? Include yourself. (range = 1-infty)
- **shift_work** (QEB31): Which of the following best describes your usual work schedule or shift at your main job). Dummy for each of the following:
  - 1: A regular daytime schedule
  - 2: A regular evening shift
  - 3: A regular night shift
  - 4: A rotating shift -- one that changes by time of day or day of week
  - 5: A split shift consisting of two distinct periods in each workday
  - 6: A flexible or variable schedule with no set hours, on call
  - 7: Some other schedule (V)
- **union:** =1 if yes; 0 otherwise
  - QEB44: Are you a member of a union OR collective bargaining unit? (1:yes, 2:no)
- **public_job:** =1 if public employee; 0 if not
  - QEB2: Are you employed by government, a private company, a non-profit organization, a single private household – OR are you self-employed or a business owner? (
    - 1: Government
    - 2: A private for-profit business
    - 3: A non-profit organization
    - 4: A single private household
    - 5: Self-employed or business owner

- **wfc_index:** Is an index constructed by taking the average response to five questions on work-family conflict, all of which are measured on the same five point Likert scale (1: very often - 5: never). The questions in the work-family conflict index are in the past three months:
  - QWF9: How often have you NOT had enough time for your family or other important people in your life because of your job?
  - QWF10: How often have you NOT had the energy to do things with your family or other important people in your life because of your job?
  - QWF11: How often has work kept you from doing as good a job at home as you could?

    - QWF12: How often have you not been in as good a mood as you would like to be at home because of your job?
    - QWF13: How often has your job kept you from concentrating on important things in your family or personal life?
- **wfc_index_d**: If wfc_index <= 2, then 1; 0 otherwise.
- **flex1** (QBP21): How hard is it for you to take time off during your work day to take care of personal or family matters -- very hard(1), somewhat hard(2), not too hard(3), or not at all hard(4)? Coded as, if >=3, then 1; 0 otherwise.
- **flex2** (QBP22B): Are you able to temporarily change your starting and quitting times on short notice when special needs arise if you check with your supervisor or manager)? "Special needs" might include having to take a car into the shop for repairs, having to take a sick child or relative to the doctor, having to meet with a teacher after school, having to stay home for a delivery, and so forth. – yes(1), no(2). Coded as, if = 1, then 1; 0 otherwise.
- **flex3** (QEB31B): Overall, how much control would you say you have in scheduling your work hours –- complete control(1), a lot(2), some(3), very little(4), or none(6)? Coded as if >=3, then 1; 0 otherwise. Coded as, if <=2, then 1; 0 otherwise.
- **flex4** (QBP22A1): Do you actually choose starting and quitting times to meet your personal needs? -- yes(1), no(2). Coded as if = 1, then 1; 0 otherwise.
- **flex5** (QBP35A): Are employees in your organization allowed to work a compressed workweek for part or all of the year? For example, can they work 10-hour days for 4 days per week instead of 8-hour days for 5 days per week OR another similar arrangement? Some employers allow compressed workweeks during the summer months, calling them "summer hours." – yes(1), no(2). Coded as if = 1, then 1; 0 otherwise.
- **Industry (indust14):**
    - agcon = if indus14 = 1 or 2, then 1; 0 otherwise.
    - mfg = if indus14 = 3, then 1; 0 otherwise.
    - trans = if indus14 = 4, then1; 0 otherwise.
    - trade = if indus14 = 5 or 6, then 1; 0 otherwise.
    - fin = if indus14 = 7, then 1; 0 otherwise.
    - serv = if indus14 >= 8 and <=13, then 1; 0 otherwise.
    - pubad = if indus14 = 14, then 1; 0 otherwise.
- **Occupation (occup7):**
    - oexec = occup7=1, then 1; 0 otherwise.
    - oprof = occup7=2, then 1; 0 otherwise.
    - otech = occup7=3, then 1; 0 otherwise.
    - osales = occup7=4, then1; 0 otherwise.
    - osupp = occup7=5, then 1; 0 otherwise.
    - oserv = occup7=6, then 1; 0 otherwise.
    - oprod= occup7=7, then 1; 0 otherwise.

References

Angrist, J and J. Pischke (2009) *Mostly harmless econometrics: An empiricist's companion*. Princeton university press: Princeton, New Jersey.

Athey, S. (2018). "The impact of machine learning on economics" *The economics of artificial intelligence: An agenda*, 507-547. University of Chicago Press.

Athey, S., and G. W. Imbens (2019). "Machine learning methods that economists should know about" *Annual Review of Economics* **11**, 685-725.

Balaj, M. (2022). "Self-reported health and the social body" *Social Theory & Health* **20**, 71-89.

Baughman, R., DiNardi D., and D. Holtz-Eakin (2003). "Productivity and wage effects of 'family-friendly' fringe benefits" *International Journal of Manpower* **24**, 247-259.

Blackmore, E., Stansfeld, S., Weller, I., Munce, S., Zagorski, B., and D. Stewart (2007). "Major depressive episodes and work stress: Results from a national population survey" *American Journal of Public Health* **97**, 2088-2093.

Bloom, N. and J. van Reenen (2006). "Management Practices, Work-Life Balance, and Productivity: A Review of Some Recent Evidence" *Oxford Review of Economic Policy* **22**, 457-482.

Boerma, T., Hosseinpoor, A. R., Verdes, E., and S. Chatterji (2016). "A global assessment of the gender gap in self-reported health with survey data from 59 countries" *BMC Public Health* **16**, 1-9.

Boodoo, U. M., Gomez, R., and M. Gunderson (2014). "Relative income, absolute income and the life satisfaction of older adults: do retirees differ from the non-retired?" *Industrial Relations Journal*, **45**, 281-299.

Boyce, C. J., Brown, G. D., and S. C. Moore (2010). "Money and happiness: Rank of income, not income, affects life satisfaction" *Psychological Science*, **21**, 471-475.

Clark, C. R., Ommerborn, M. J., Moran, K., Brooks, K., Haas, J., Bates, D. W., and A. Wright (2021). "Predicting self-rated health across the life course: health equity insights from machine learning models" *Journal of General Internal Medicine* **36**, 1181-1188.

Cotti, C, Haley, M, and L. Miller (2014) "Workplace flexibilities, job satisfaction and union membership in the US workforce" *British Journal of Industrial Relations* **52**, 403-425.

Cotti, C, Haley, M, and L. Miller (2017) "Assessing the impact of different workplace flexibilities on workplace stress in the presence of varying degrees of job control" *Applied Economics Letters* **24**, 198-201.

Dunn, J., Landry, S., and K. Binsted (2021). "Measuring the impact of stressors through self-reporting on the temporal nature of how perceived stress emerges and dissipates" *Journal of Mental Health Disorders* **1**, 1-9.

Erhel, C., Guergoat-Larivière, M., and M. Mofakhami (2024). "Diversity of flexible working time arrangements and workers' health: An analysis of a workers' panel and linked employer-employee data for France" *Social Science & Medicine* **356**, 117129.

Fischer, J. A., and A. Sousa-Poza (2009). "Does job satisfaction improve the health of workers? New evidence using panel data and objective measures of health" *Health Economics* **18**, 71-89.

Gandhi, K., Lim, E., Davis, J., and J. J. Chen (2020). "Racial-ethnic disparities in self-reported health status among US adults adjusted for sociodemographics and multimorbidities, National Health and Nutrition Examination Survey 2011–2014" *Ethnicity & Health* **25**, 65-78.

Gevaert, J., Van Aerden, K., De Moortel, D., and C. Vanroelen (2021). "Employment quality as a health determinant: Empirical evidence for the waged and self-employed" *Work and Occupations* **48**, 146-183.

Grzywacz, J., Casey, P., and F. Jones (2007). "Workplace flexibility and employee health behaviors: A cross-sectional and longitudinal analysis" *Journal of Occupational & Environmental Medicine* **49**, 1302-1309.

Gumà-Lao, J., and B. Arpino (2023). "A machine learning approach to determine the influence of specific health conditions on self-rated health across education groups" *BMC Public Health* **23**, 131.

Haley, M and L. Miller (2015) "Correlates of flexible working arrangements, stress, and sleep difficulties in the US workforce: does the flexibility of the flexibility matter?" *Empirical Economics* **48**, 1395-1418.

Haley, M, and L. Miller (2023). "Predicting preferences for flexible working arrangements in future employment: A gender analysis" *Economics Bulletin* **43**, 882-893.

Halpern, D. (2005). "How time-flexible work policies can reduce stress, improve health, and save money" *Stress and Health* **21**, 157-168.

Harnois, C. E., and J. L. Bastos (2018). "Discrimination, harassment, and gendered health inequalities: do perceptions of workplace mistreatment contribute to the gender gap in self-reported health?" *Journal of Health and Social Behavior* **59**, 283-299.

Hosseinpoor, A. R., Stewart Williams, J., Amin, A., Araujo de Carvalho, I., Beard, J., Boerma, T., ... and S. Chatterji (2012). "Social determinants of self-reported health in women and men: understanding the role of gender in population health" *PloS One*, **7**, e34799.

Ishida, M., Montagni, I., Matsuzaki, K., Shimamoto, T., Cariou, T., Kawamura, T., ... and T. Iwami (2020). "The association between depressive symptoms and self-rated health among university students: a cross-sectional study in France and Japan" *BMC Psychiatry* **20**, 1-10.

Johansson, E., Böckerman, P., and A. Lundqvist (2020). "Self-reported health versus biomarkers: does unemployment lead to worse health?" *Public Health* **179**, 127-134.

Joyce, K., Pabayo, R., Critchley, J. A., and C. Bambra (2010). "Flexible working conditions and their effects on employee health and wellbeing" *Cochrane Database of Systematic Reviews* **2**.

Kim, J, Henly, J, Golden, L, and S. Lambert (2020) "Workplace flexibility and worker well-being by gender" *Journal of Marriage and Family* **82**, 892-910.

Kim, E. S., Delaney, S. W., Tay, L., Chen, Y., Diener, E. D., and T. J. Vanderweele (2021). "Life satisfaction and subsequent physical, behavioral, and psychosocial health in older adults" *The Milbank Quarterly* **99**, 209-239.

Kino, S., Hsu, Y. T., Shiba, K., Chien, Y. S., Mita, C., Kawachi, I., and A. Daoud (2021). "A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects" *SSM-population Health* **15**, 100836.

Koivumaa-Honkanen, H., Honkanen, R., Viinamäki, H., Heikkilä, K., Kaprio, J., and M. Koskenvuo (2000). "Self-reported life satisfaction and 20-year mortality in healthy Finnish adults" *American Journal of Epidemiology* **152**, 983-991.

Kööts–Ausmees, L., and A. Realo (2015). "The association between life satisfaction and self–reported health status in Europe" *European Journal of Personality* **29**, 647-657.

Kudrnáčová, M., and A. Kudrnáč (2023). "Better sleep, better life? Testing the role of sleep on quality of life" *PLoS One **18***, e0282085.

Lee, S. W., Choi, J. S., and M. Lee (2020). "Life satisfaction and depression in the oldest old: a longitudinal study" *The International Journal of Aging and Human Development* **91**, 37-59.

Lin, M. H., Chen, L. J., Huang, S. T., Meng, L. C., Lee, W. J., Peng, L. N., ... and L. K. Chen (2022). "Age and sex differences in associations between self-reported health, physical function, mental function and mortality" *Archives of Gerontology and Geriatrics* **98**, 104537.

Lindström, M. (2009). "Marital status, social capital, material conditions and self-rated health: a population-based study" *Health Policy* **93**, 172-179.

Mahajan, S., Caraballo, C., Lu, Y., Valero-Elizondo, J., Massey, D., Annapureddy, A. R., ... and H. M. Krumholz (2021). "Trends in differences in health status and health care access and affordability by race and ethnicity in the United States, 1999-2018" *Jama* **326**, 637-648.

McNeil Smith, S., Williamson, L. D., Branch, H., and F. D. Fincham (2020). "Racial discrimination, racism-specific support, and self-reported health among African American couples" *Journal of Social and Personal Relationships* **37**, 779-799.

Mensah, A., and N. K. Adjei (2020). "Work-life balance and self-reported health among working adults in Europe: a gender and welfare state regime comparative analysis" *BMC Public Health* **20**, 1-14.

Molina, T. (2016). "Reporting heterogeneity and health disparities across gender and education levels: Evidence from four countries" *Demography* **53**, 295-323.

Olsson, M., Currow, D. C., and M. P. Ekström (2022). "Exploring the most important factors related to self-perceived health among older men in Sweden: a cross-sectional study using machine learning" *BMJ open* **12**, e061242.

Pollmann-Schult, M. (2018). "Parenthood and life satisfaction in Europe: The role of family policies and working time flexibility" *European Journal of Population* **34**, 387-411.

Qin, F. Y., Lv, Z. Q., Wang, D. N., Hu, B., and C. Wu (2020). Health status prediction for the elderly based on machine learning. *Archives of Gerontology and Geriatrics* **90**, 104121.

Roxo, L., Bambra, C., and J. Perelman (2021). "Gender equality and gender inequalities in self-reported health: a longitudinal study of 27 European countries 2004 to 2016" *International Journal of Health Services* **51**, 146-154.

Subramanian, S. V., Huijts, T., and M. Avendano (2010). "Self-reported health assessments in the 2002 World Health Survey: how do they correlate with education?" *Bulletin of the World Health Organization* **88**, 131-138.

Svensen, E., Arnetz, B. B., Ursin, H., and H. R. Eriksen (2007). "Health complaints and satisfied with the job? A cross-sectional study on work environment, job satisfaction, and subjective health complaints" *Journal of Occupational and Environmental Medicine* **49**, 568-573.

Vaingankar, J. A., Chong, S. A., Abdin, E., Siva Kumar, F. D., Chua, B. Y., Sambasivam, R., ... and M. Subramaniam (2020). "Understanding the relationships between mental disorders, self-reported health outcomes and positive mental health: findings from a national survey" *Health and Quality of Life Outcomes* **18**, 1-10.

Varian, H. R. (2014). "Big data: New tricks for econometrics" *Journal of Economic Perspectives* **28**, 3-28.

Wallace, M. L., Buysse, D. J., Redline, S., Stone, K. L., Ensrud, K., Leng, Y., ... and M. H. Hall (2019). "Multidimensional sleep and mortality in older adults: a machine-learning comparison with other risk factors" *The Journals of Gerontology: Series A* **74**, 1903-1909.

Whiting, R., and Bartle-Haring, S. (2022). “Variations in the association between education and self-reported health by race/ethnicity and structural racism” *SSM-Population Health* **19**, 101136.

Wu, S., Wang, R., Zhao, Y., Ma, X., Wu, M., Yan, X., and He, J. (2013). “The relationship between self-rated health and objective health status: a population-based study” *BMC Public Health* **13**, 1-9.

Zella, S. (2017). “Marital status transitions and self-reported health among Canadians: A life course perspective” *Applied Research in Quality of Life* **12**, 303-325.