Predictors of Mental Health Among U.S. Adult Population

Author: Dung Pham

December, 2021

The code to reproduce this report is available on Github:
https://github.com/dphamvu/NHANES

Executive Summary

Problem

To date, most studies have examined predictors of mental health only among a specific subset of the population. Given the high prevalence of mental health problems among Americans and its profound consequences at both the individual and societal levels, I seek to investigate various predictors of mental health risks among the US adult population.

Data

Data was obtained from the National Health and Nutrition Examination (NHANES) Survey 2017-2020 by the Center for Disease Control and Prevention. NHANES survey combines interviews and physical examinations to assess the health and nutritional status of adults and children in the United States. The explanatory variables come from 10 different areas: demographics, mental health/depression screening questions, physical activity, smoking, sleep, medical history, diabetes, blood pressure and cholesterol, occupation, and weight history. There are multiple items within each group of variables for a total of 41 features included in the analysis. The response variable (mental health/depression score) was created by summing up all ten items in the mental health/depression screening dataset. Each item has a value ranging from 0 to 3. Therefore, the response variable is on a scale from 0 to 30, with greater value indicative of higher mental health risks.

Analysis

Missing values were removed from the dataset, resulting in a total of 922 responses in the final data file. Data was split into training and test sets for a 80-20 ratio. First, I explored the relationships between different variables on the training data. Six different cross-validated predictive models were then built, including ordinary least squares, ridge regression, LASSO regression, elastic net regression, random forest, and boosting. For the tree-based models, the boosted model had the lowest test error as well as the lowest test error among the six methods. Within the regression models, OLS had the lowest test error, followed by ridge and elastic net regression (which had the same predictive performance).

Conclusions

All six models consider sleep and ratio of family income to poverty as important predictors of mental health risks. Gender, vigorous-intensity recreational activities and self-perception about weight are also shared by several predictive models. These findings hope to inform initiatives and policies to address the sleep deprivation epidemic among Americans, provide more support for women and low-income populations, allocate more resources for mental wellness, physical and recreational activities, as well as raise awareness about positive body image. I concluded by discussing several limitations and providing recommendations for future research directions.

Introduction

Background information

Mental health is a global public health issue[1]. According to a recent study by Nochaiwong et al. (2021), the estimated global prevalence of mental health problems are widespread: 50.0% for psychological distress; 36.5% for stress; 28.0% for depression; 27.6% for sleep problems; 26.9% for anxiety; and 24.1% for post-traumatic stress symptoms. COVID-19 has exacerbated mental health outcomes among different populations around the world (Nochaiwong et al., 2021; Wu et al., 2021).

According to the World Health Organization[2], depression is the leading cause of disability worldwide and contributes significantly to the global burden of disease. The World Economic Forum[3] cited a Lancet commission report that estimated the cost to the global economy of all mental health problems to amount to $16 trillion by 2030. Mental health challenges was also featured on the agenda at the World Economic Forums' Annual Meeting in 2019 in Davos.

At the individual level, mental health plays a pivotal role at every stage of life, from childhood and adolescence through adulthood and old age. It affects how we think, feel, act, relate to others and make choices[4]. Mental health has a significant impact on physical health. For example, depression increases the risk for chronic conditions such as diabetes[5], heart disease[6] and stroke[7]. Mental illness also impacts social determinants of health such as homelessness, school dropout, marital instability, and economic insecurity (Corrigan et al., 2012; Ljungqvist et al., 2016; Hjorth et al., 2016).

Given the prevalence and influence of mental health in major life outcomes at both the individual and societal levels, it is imperative to understand the predictors of mental well-being for early detection and prevention. A number of studies have set out to identify predictors of mental health among a specific subset of populations. Female gender was associated with risk of PTSD among health workers (Hennein et al., 2021). Predictors of mental health status among older United States adults with pain included physical health status, employment, and education higher than high school (Axon & Chien, 2021). In a study conducted during the COVID-19 pandemic among Iranian parents of children with cerebral palsy, being married, low educational level and low income were significantly related to anxiety while physical problem was significantly correlated with depression (Farajzadeh et al., 2021). Marital status was also a major predictor of workers' mental health and fatigue levels (Alroomi & Mohamed, 2021). Trabelsi et al (2020) found sleep quality and physical activity to be predictors of mental

[1] Center for Disease Control and Prevention (2020). https://www.cdc.gov/pcd/collections/Mental_Health_Is_a_Global_Public_Health_Issue.htm
[2] World Health Organization. (n.d.) https://www.who.int/news-room/fact-sheets/detail/depression
[3] World Economic Forum (2019). https://www.weforum.org/agenda/2019/01/this-is-the-worlds-biggest-mental-health-problem/
[4] Center for Disease Control and Prevention. (n.d.). https://www.cdc.gov/mentalhealth/learn/index.htm).
[5] Center for Disease Control and Prevention. (n.d.). https://www.cdc.gov/diabetes/managing/mental-health.html
[6] Center for Disease Control and Prevention. (n.d.). https://www.cdc.gov/heartdisease/mentalhealth.htm
[7] Center for Disease Control and Prevention. (n.d.). https://www.cdc.gov/mentalhealth/learn/index.htm

well-being among older adults in different regions of the world during COVID-19 lockdown. Among Canadian freshman college students, social support, sleep quality and exercise frequency at entry to university were associated with positive mental health screening measured at the end of their first year (Duffy et al., 2020). Research conducted among Danish adults (aged 16 and above) revealed that lower SES standing (education, income and employment status) was associated with increased likelihood of low mental well-being and common mental disorders, but did not significantly predict high mental well-being (Santini et al., 2020). Relational/recreational behaviors (informal and formal social participation, social support and recreational activity), on the other hand, were associated with positive mental well-being (Ibid.).

Despite the plethora of research interest in mental health, few studies have set out to systematically investigate predictors of mental well-being among the general population in the US, especially factors related to health behaviors and lifestyles rather than trait personality.

Analysis goal

This report seeks to examine potential predictive factors of mental health risks among US adult populations (age 20 and above). The features come from a variety of areas, including health behaviors and lifestyles such as smoking status, sleep, physical activity, self-perception about one's weight, along with demographic characteristics, physical health status/medical history and occupation/working hours. These variables will be assessed on their association with the response variable — mental health/depression screening scores. Six different regression and tree-based methods will be tested to identify the important predictor variables based on the lowest test root mean squared error (test RMSE), using the intercept-only model as the benchmark.

Significance

Mental health issues are prevalent in the US. In 2019, approximately one in five (51 million) U.S. adults aged 18 years or above reported any mental illness[8]. This report hopes to contribute to fill in a small gap in the literature on factors that can predict mental health risks among the US general population, thus to identify mechanisms to increase mental wellness for communities at large, especially amid the aggravating impacts of COVID-19.

[8] (Substance Abuse and Mental Health Services Administration. (2020). Key substance use and mental health indicators in the United States: Results from the 2019 National Survey on Drug Use and Health (HHS Publication No. PEP20-07-01-001). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. Retrieved from https://www.samhsa.gov/data/sites/default/files/reports/rpt29393/2019NSDUHFFRPDFWHTML/2019NSDUHFFR1PDFW090120.pdf.).

Data

Data source

The data was obtained from CDC's National Health and Nutrition Examination (NHANES) survey. NHANES is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey examines a nationally representative sample of about 5,000 persons each year located in counties across the country, 15 of which are visited each year. The survey combines both interviews and physical examinations. The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel. Due to the impacts of COVID-19, data for the cycle 2019-2020 was only collected up until March 2020 and therefore was not nationally representative. The CDC thus combined this data with the 2017-2018 data to form the current dataset. Thirty five different sets of data were manually downloaded and carefully examined, ten of which were merged to form the final dataset for analysis.

Link to the online data can be accessed at:
https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?cycle=2017-2020

Data cleaning

All 35 available datasets were downloaded and merged together. However, after the removal of missing (NA) values, the dataset did not meet the requirements of having at least 500 observations. I thus drew on the literature to select 10 areas of interest to form the final dataset, including demographic characteristics, mental health/depression screening questions, physical activity, smoking, sleep, medical history, diabetes, blood pressure and cholesterol, occupation, and weight history. Each area contains multiple question items. Being mindful not to cherry-pick features, I tried to, whenever possible, include all items within each area. However, this was not always feasible due to insufficient data quality. After thoroughly familiarizing myself with the codebook, I examined the data and found that the researchers were not consistent in coding the "Don't know" (DK) and "Refused" (RF) answers (different numeric values such as 7, 9, 77, 99, 777, 999, 7777, 9999, 77777, 99999 etc. were used). I thus went through each dataset to replace the respective values with N/A while making sure that the real numeric values (not associated with DK and RF) remained intact.

Each of the 10 datasets was cleaned separately. I first renamed the variables to facilitate later analysis, selected the features within each dataset, merged all datasets, dropped all N/A values and saved the final dataset to a .csv file. I originally attempted to mutate categorical variables to dummy variables. However, to ensure the nuances of different levels of the variables would not be missed, I retained the levels of each categorical variable in its original format. It is also

worth noting that only participants aged 20 and above were included in the analysis as some demographic information (i.e., education, marital status) was not available for those under 20.

The response variable was created by summing up all ten items in the mental health/depression screening dataset. Each item has a value ranging from 0 to 3. Thus, the response variable is on a scale from 0 to 30, with higher value indicative of higher mental health risks.

Data description

Observations

There are 922 observations in the data. Each observation represents an individual who was interviewed and examined during the NHANES 2017-March 2020 survey period.

Response Variable

The response variable — mental health/depression screening score is continuous. This item was created by summing up all ten items within the mental health/depression screening questionnaire. The screening instrument was adapted from the Patient Health Questionnaire to assess the frequency of depression symptoms over the past two weeks. For each symptom question, points ranging from 0 to 3, are associated with the response categories "not at all," "several days," "more than half the days," and "nearly every day." This screening questionnaire incorporates DSM-IV depression diagnostic criteria on several areas: interest, negative mood/feelings/thoughts, trouble with sleep, energy level, appetite, concentration, speed of mobility and speech, and difficulty these problems have caused. The response variable is thus on a scale of 0 to 30, with higher score suggestive of greater mental health/depression risks.

Explanatory Variables

Forty-one features were present in the data. Most of them are categorical variables; only a few are continuous. The features come from 10 different groups: demographic characteristics, mental health/depression screening questions, physical activity, smoking, sleep, medical history, diabetes, blood pressure and cholesterol, occupation, and weight history (please see Appendix for a detailed specification of the variables).

Data allocation

The observations in the dataset were divided into two subsets: one training dataset for predictive model building and a test dataset for model evaluation. I used an 80-20 split in which the training dataset contained 80% of the observations and the test set consisted of the remaining observations (20%). A random seed was also set to ensure that each train-test split for each class of method would lead to reproducible results.

Data exploration

To provide an overview of the data, I explored summary statistics on the training dataset (n=738). The response variable (mental health screening scores) has a mean of 4.9 on a scale from 0 to 30, with higher score indicative of higher mental health risks. The histogram in Figure 1 suggests that the distribution of the response variable is right-skewed, with the median value of 3. The median age of the sample is 44 with non-Hispanic White (43%) and non-Hispanic Black (22%) as the majority ethnicity/racial groups. Nearly half of the sample is married or living with a partner (57%); 23% reported to be single. The majority of participants were US-born (80%) and have some college or AA degree (41%). About 18% of the sample have obtained a college degree or higher. The gender breakdown in the training data is 41% and 59% for female and male participants respectively. Based on the US Census definition, 17% of the sample reported to be below the poverty level (having an income-to-poverty ratio of less than 1). The median ratio of family income to poverty in the training dataset is 2.32 (232% above poverty baseline), and 15% of subjects are at least 500% above the poverty threshold (having a ratio of 5 and above).

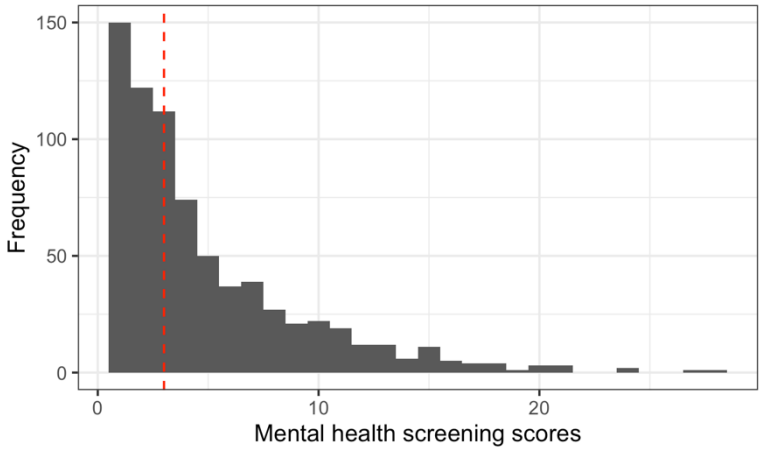
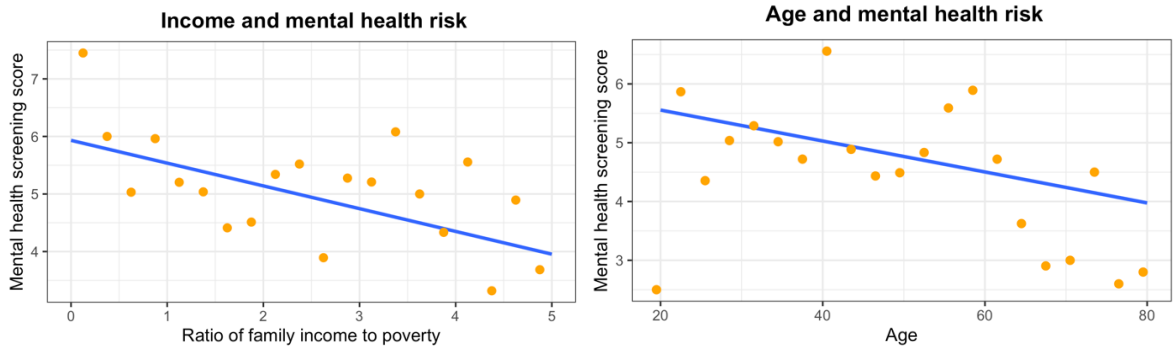
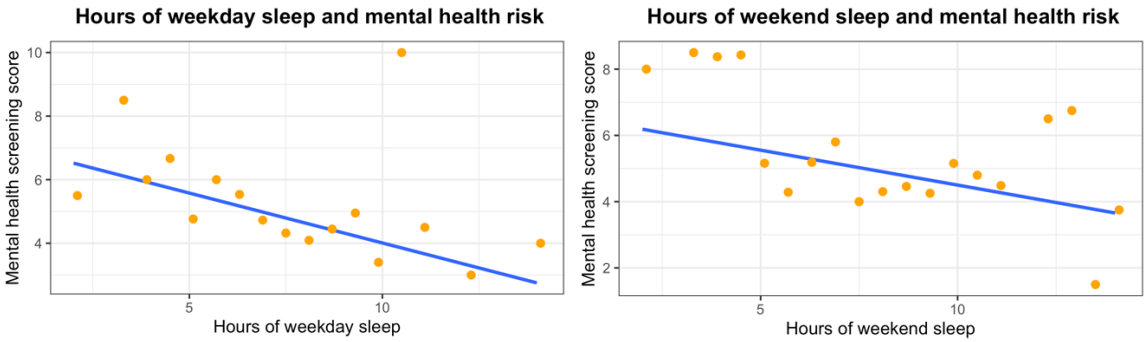


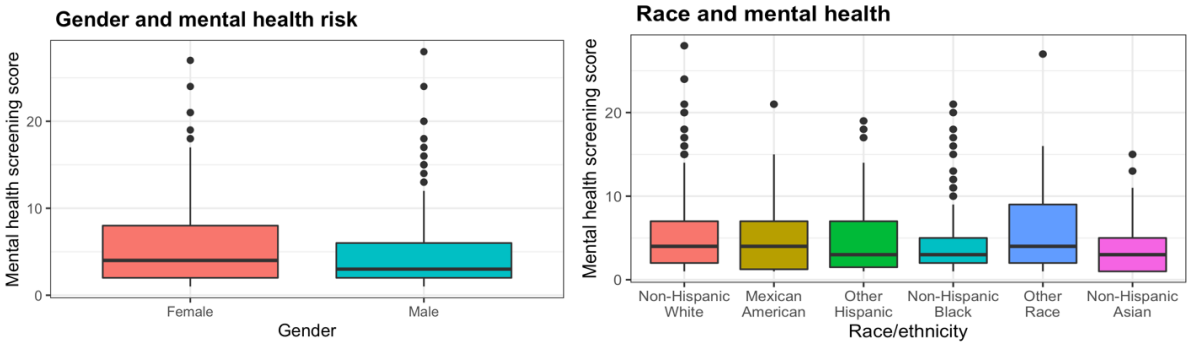
Figure 1. Histogram of the response variable (mental health screening scores; vertical dashed red line indicates the median)

Ratio of family income to poverty, hours of weekday sleep, hours of weekend sleep, age tend to have negative correlation with mental health risks.

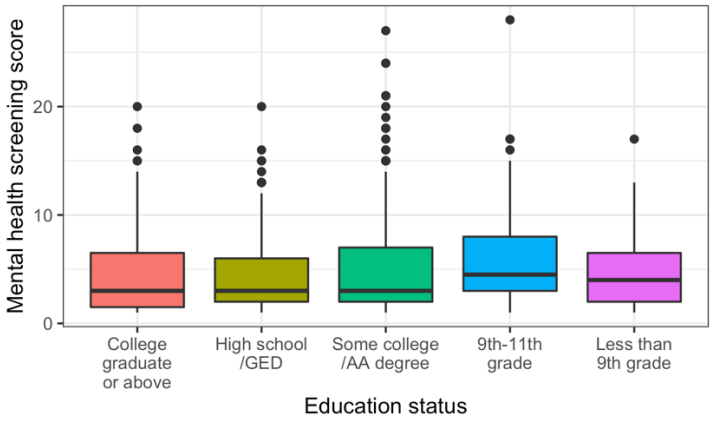




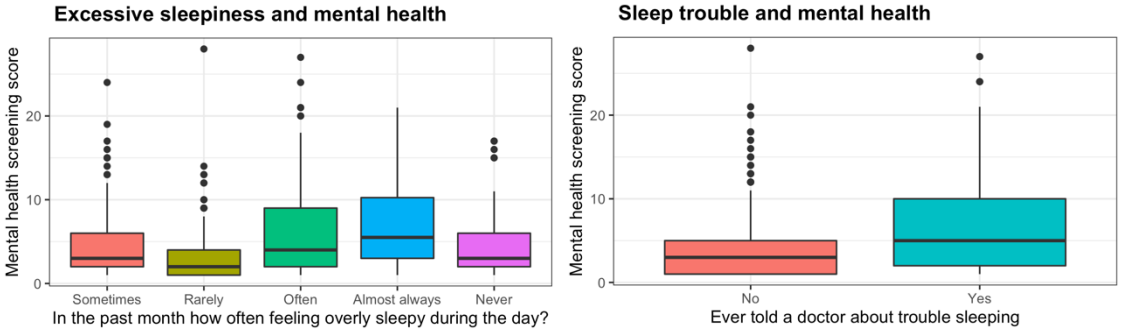
Two boxplots below suggest a potential relationship between gender and mental health in which females appear to have higher mental health screening scores than males. There also seems to be differences in mental health risks among different racial/ethnicity groups. Non-Hispanic Blacks, Non-Hispanic Asian and Other Hispanic tend to score on the lower spectrum of the mental health screening risk.



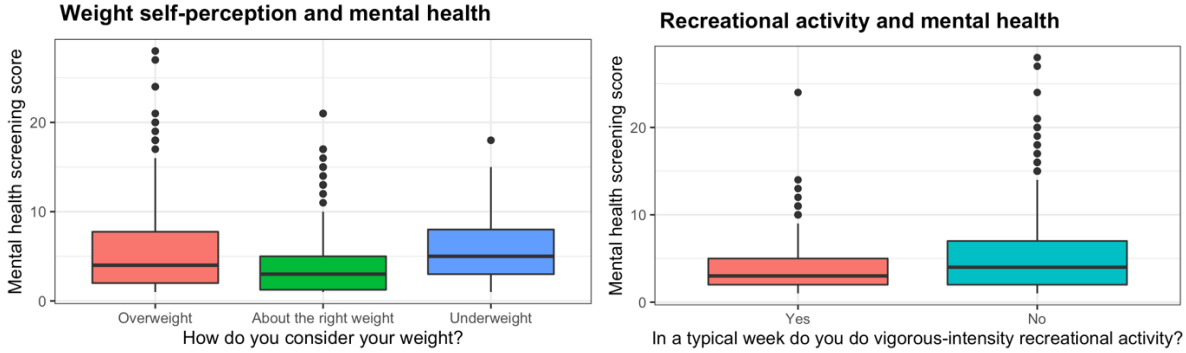
Those who completed 9th to 11th grade (or completed 12th grade with no diploma) seemed to fare significantly worse on mental health outcomes compared to people with other educational status. Generally, the boxplot seems to suggest that lower education is associated with higher risk of depression.



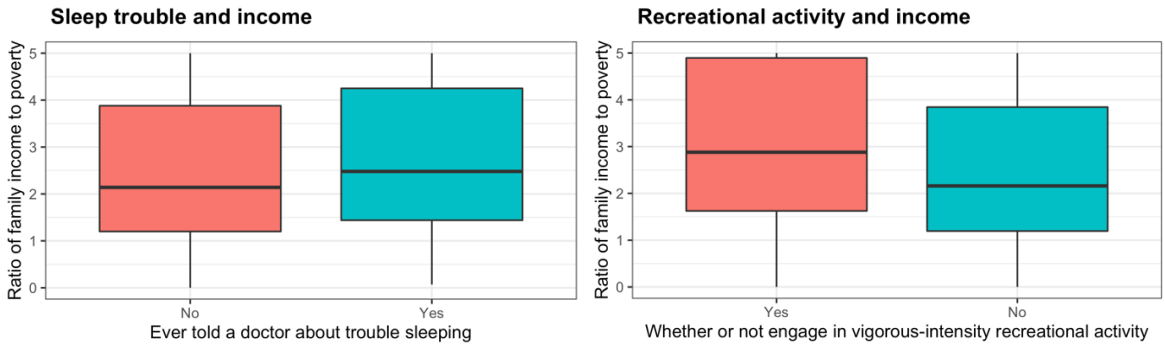
Exploratory data analysis on sleep and mental health suggests a potential relationship between the two variables. Those who ever told a doctor about their sleep troubles tend to score higher on the mental health screening items than their counterparts. Individuals who often and almost always feel overly sleepy during the day seem to be at higher risk of mental health problems than those who experience excessive sleepiness less frequently.



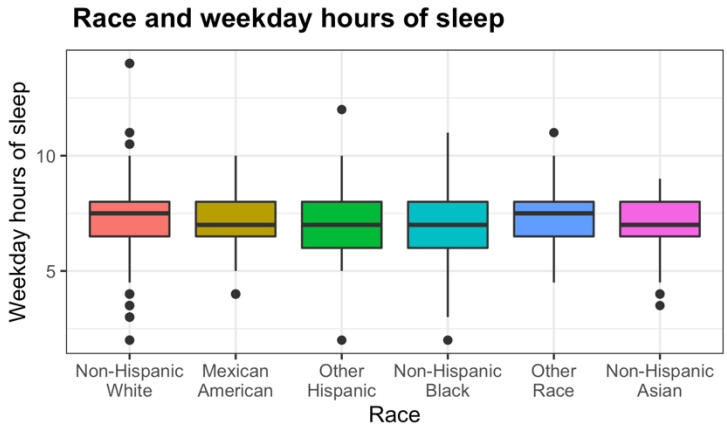
Data exploration shows that individuals who engage in vigorous-intensity recreational activity in a typical week tend to report better mental health outcomes. Intriguingly, the boxplot below seems to suggest that those who consider themselves to be underweight fare worse on self-reported mental health symptoms than those who regardless themselves as overweight or about the right weight.



In examining the potential covariation between ratio of family income to poverty and other variables, the following boxplots offer interesting insights. Higher ratio of family income to poverty seems to be correlated with higher likelihood of engaging in vigorous-intensity recreational activities in a typical week. However, those with higher incomes also more likely tend to have ever told their doctor about their trouble with sleep.



Last but not least, exploratory analysis on race and weekday hours of sleep provide further food for thoughts. In the previous boxplot on the relationship between race and mental health, Non-Hispanic Blacks, Non-Hispanic Asian and Other Hispanic appear to be at lower risk of mental health problems. The boxplot below shows that these three groups in addition to Mexican Americans tend to have fewer number of sleeps during weekday compared to Non-Hispanic Whites and other races. Though having about similar number of sleep hours on the weekday, Mexican Americans reported worse mental well-being.



Though exploratory analysis was done on all 41 features in relation to the response variable, only a selected few of notable results were presented in this section. Such preliminary exploration reveal potential relationships between mental health risks and a number of demographic characteristics, sleep, self-perception about weight and recreational activities in a typical week. However, there are still contradictory insights that suggest complexity and nuances in the variations and interactions between the response and features, which calls for different predictive methods in the following section to shed light on.

Modeling

Regression Methods

Ordinary Least Squares Regression (OLS)

First, I conducted an ordinary least squares (OLS) regression of mental health scores on all 41 features. The OLS regression showed the following variable responses to be statistically significant at the .05 level: female gender (default comparison: male); ratio of family income to poverty; high school graduate/GED or equivalent (default comparison: college graduate or above); ever being told by a doctor to have a liver condition (default comparison: not having a liver condition), consider oneself at about the right weight (default comparison: overweight). For categorical variables, significant signal only means that a certain level within a categorical variable is significantly different from its default level. For example, female is significantly different from male in mental health scores as male serves as the default level of the gender variable.

At the .01 level, the following variables showed statistical significance: often feeling overly sleepy during the day – 5 to 15 times a month (default comparison: sometimes – 2 to 4 times a month); not engaging in vigorous-intensity recreational activities (default comparison: engaging in vigorous-intensity recreational activities).

At the .001 level, mental health scores among those who ever told a doctor about their trouble with sleep were significantly higher than those who did not. Those who almost always feel overly sleepy during daytime (16-30 times a month) also scored significantly higher on mental health risk screening than those who reported only sometimes feel excessive sleepiness (2-4 times a month).

The multiple R-squared indicates that these features account for 23.5% of the variation in the response variable. After adjusting for the large number of predictor variables in the model, the adjusted R-squared value decreases to only 0.1725 (17.25%).

Penalized regression

To address potential high variance in the OLS regression model with a large number of explanatory variables, I employed cross-validated ridge, LASSO (Least Absolute Shrinkage and Selection Operator) and elastic net regression, choosing lambda based on the one-standard-error rule. Elastic net regression contains all the features selected by LASSO, in addition to 22 other features. Below is the list of the variables selected by each method.

- **LASSO**: ever told your doctor/health profession that you had trouble sleeping, ratio of family income to poverty, how often feel overly sleepy during the day.

- **Elastic net regression**: age, high blood pressure and cholesterol, ever told your doctor/health profession that you had trouble sleeping, education, gender, moderate-intensity activity at work, now increasing exercise, now controlling or losing weight, now reducing fat in diet, now reducing salt in diet, now smoking, race, ratio family income to poverty, hours of weekday sleep, hours of weekend sleep, how often feel overly sleepy during the day, told by doctor to have COPD, told by doctor to exercise, told by doctor to have a liver condition, told by doctor to have a stroke, told by doctor to have thyroid problem, told by doctor to be overweight, doing vigorous-intensity recreational activities, want to weigh more or less or the same, self-perception about one's weight.

For ridge regression, the CV plot is shown in Figure 2 below. Figure 3 displays the trace plot with top 10 features.

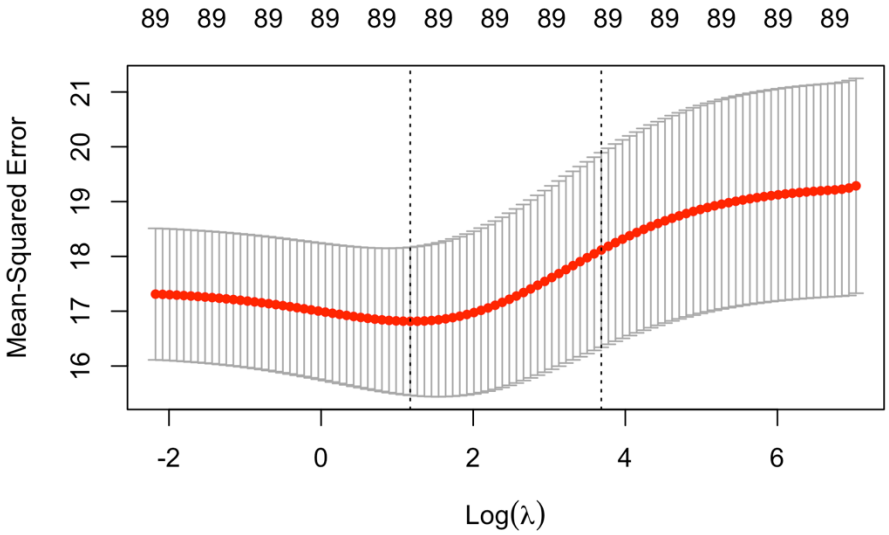


Figure 2. Ridge CV plot

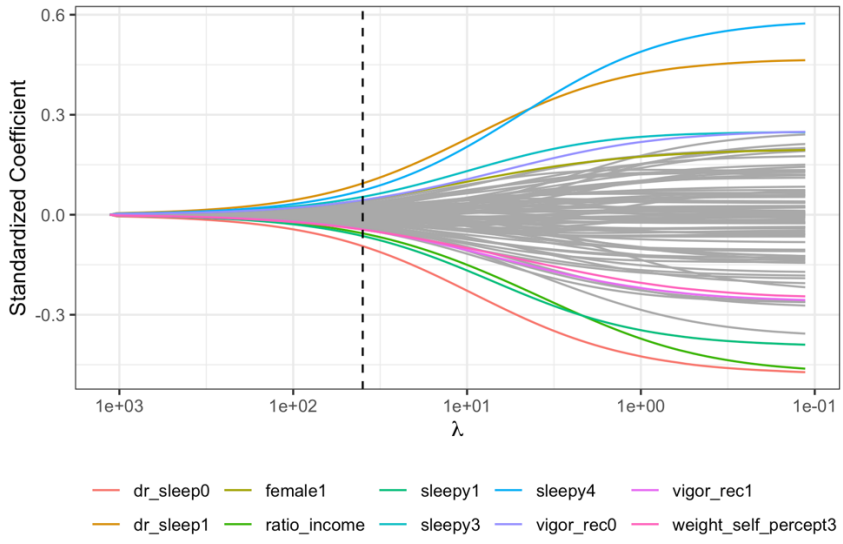


Figure 3. Ridge trace plot

For LASSO, Figure 4 shows the CV plot while Figure 5 illustrates the trace plot. The six selected features and their coefficients in the LASSO model can be found in Table 1.

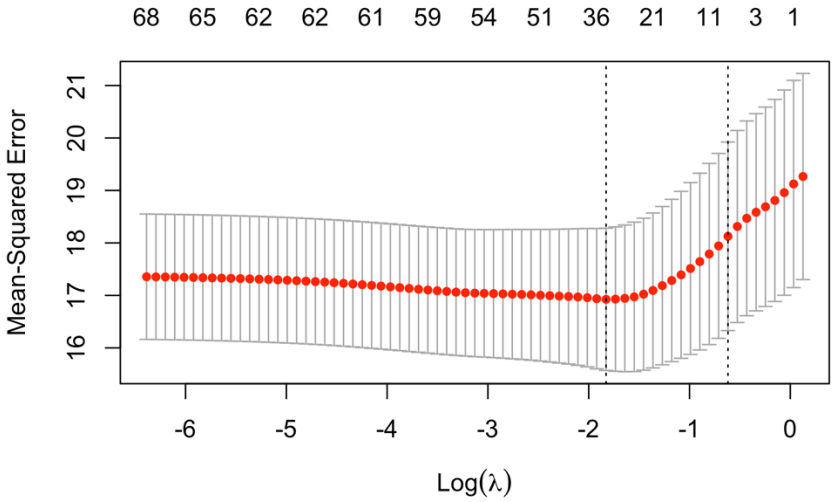


Figure 4. LASSO CV plot

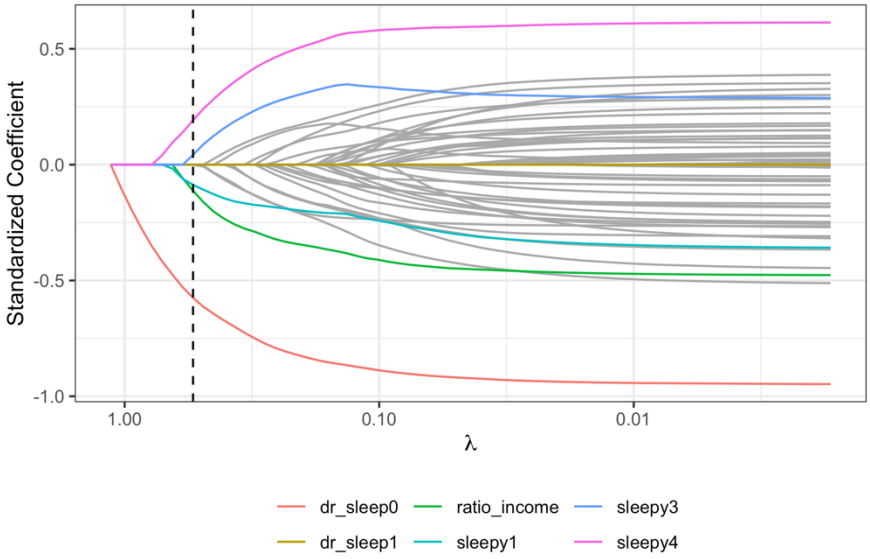


Figure 5. LASSO trace plot

Table 1. LASSO's selected features and their coefficients

LASSO features	**Coefficient**
dr_sleep0	-0.5745
sleepy4	0.1916
ratio_income	-0.1125
sleepy1	-0.0871
sleepy3	0.0420
dr_sleep1	1.1885e-13

For elastic net regression, Figure 6 shows the CV plot and Figure 7 refers to the trace plot with top 10 selected features. Table 2 details coefficients of the 10 most important features.

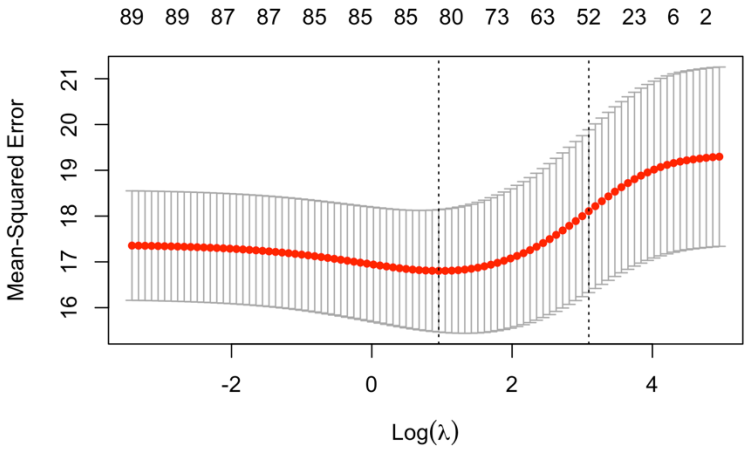


Figure 6. Elastic net regression CV plot

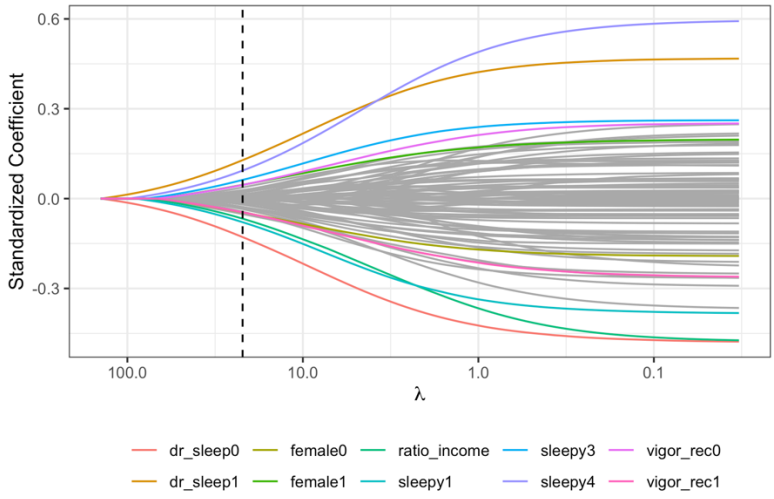


Figure 7. Elastic net regression trace plot

Table 2. Elastic net regression: Top 10 features and their coefficients

Elastic net regression features	Coefficients
dr_sleep0	-0.1283
dr_sleep1	0.1283
sleepy4	0.0935
sleepy1	-0.078
ratio_income	-0.0669
sleepy3	0.0623
weight_self_percept3	-0.0515
female1	0.0460
female0	-0.0460
vigor_rec0	0.04558

Tree-based methods

Random forest

Random forest is one of the state-of-the-art tools for prediction. For best predictive performance, *m* (the number of variables to sample at each split) was tuned via out-of-bag error (OOB). The model was trained on different values of *m*, ranging from 1 to 41 as the total number of features in the data was 41. Figure 8 displays that OOB error was the lowest at *m* value of 5. *B* parameter (the number of bootstrap samples) was set to the default value of 500. The random forest model was then fitted using the tuned *m* and *B* values. To confirm that the OOB error had indeed flattened out, I plotted a graph of OOB errors using the tuned *m* value (Figure 9).

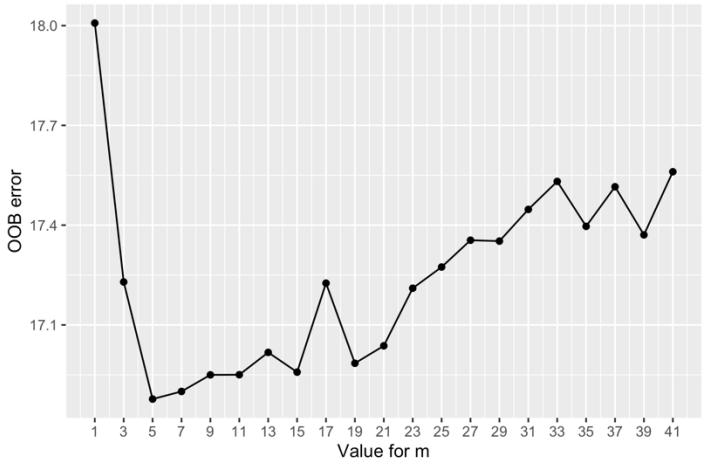


Figure 8. OOB error vs. *m*

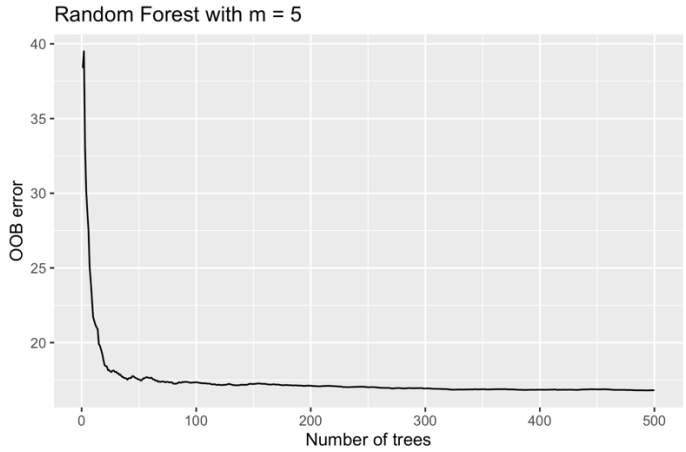


Figure 9. OOB error vs. number of trees, using *m* value of 5

Random forest provides two measures of variable importance. Mean Decrease Accuracy (%IncMSE) shows how much the model accuracy will decrease if we exclude a certain variable. Mean Decrease Gini (IncNodePurity) measures variable importance based on the Gini impurity index, indicating the total decrease in node impurity that results from splits over a given variable averaged over all trees. The higher the value of mean decrease accuracy or mean decrease Gini, the higher the importance of the variable to the model. Based on Figure 10, how often an individual feel overly sleepy during the day is an important variable in both measures. Hours of weekday sleep and race are also among the top prominent variables.

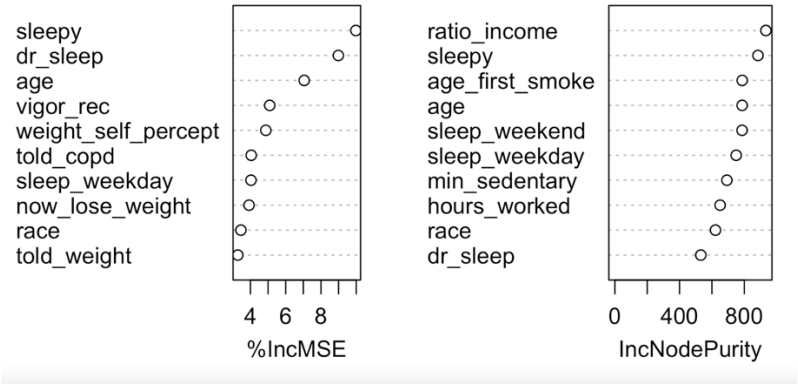


Figure 10. Random Forest Variable Importance Measures

Boosting

Boosting is another method that aggregates multiple decision trees to reduce variance and improve prediction. I tested out with the default parameters of 100 trees, an interaction depth of 1, a shrinkage factor of 0.1, and a subsampling fraction π of 0.5. Unlike random forests, the number of trees *B* controls the complexity of the fit and therefore must be tuned via cross-validation. After experimentation with different shrinkage factors and number of trees, the optimal number of trees was identified to be 57. Using this number, I tuned the interaction depth of the model by testing different interaction depths from 1 to 5. Figure 11 shows that an

interaction depth of 4 results in the smallest cross-validation error with 34 trees. Table 3 displays the top 10 most important variables selected by the boosted model.

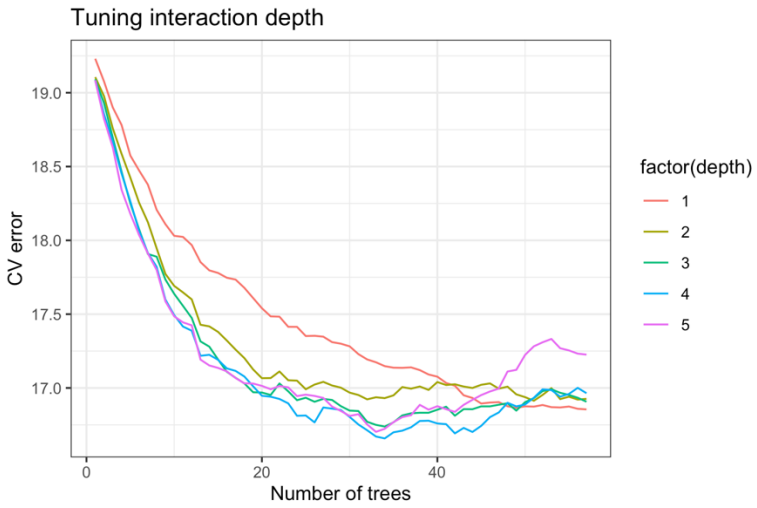


Figure 11. Tuning interaction depth

Table 3. Top 10 features in the boosted model and their relative importance

Variable	Relative influence
How often feeling overly sleepy during the day	16.561
Ever told a doctor about trouble sleeping	11.88
Age	11.441
Age started smoking regularly	8.567
Ratio of family income to poverty	8.012
Hours of weekend sleep	7.255
Hours of weekday sleep	5.932
Race	5.108
Education	3.482
Gender	2.434

Figure 12 displays the partial dependence plots for ratio of family income to poverty and how often one feels overly sleepy during the day. These are top variables in the boosted model. In the first partial dependence plot, we observe a sharp decrease in mental health screening score when the ratio income moves from 0.5 to 1. The graph then slowly levels off when ratio income reaches around 4.2. The partial dependence plot for feeling of excessive sleepiness during the day suggests that those who report often or almost always experiencing drowsiness tend to score the highest on mental health screening scores. The graph implies a general positive relationship between mental health risks and frequency of excessive sleepiness.

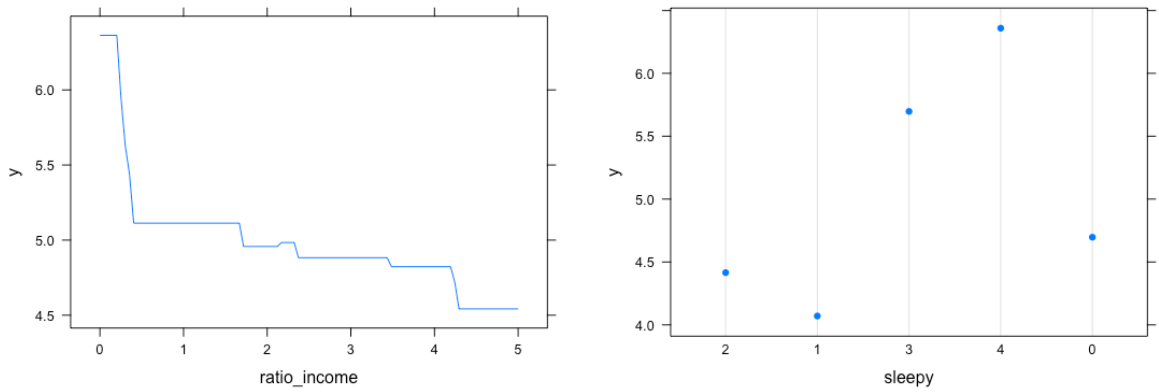


Figure 12. Partial dependence plot: Ratio income and how often one feels overly sleepy during the day

Conclusions

Method comparison

Table 4 shows the test root mean squared error (test RMSE) for all the methods considered. Known for its high predictive accuracy, the boosted model results in the lowest test RMSE of 4.00, closely followed by ordinary least squares regression (4.095). It's worth noting that the latter method (OLS) has an adjusted R-squared value of only 0.172 with a lot of features (41 features) in its model. Ridge, elastic net regression and LASSO also performed not too far below boosting and OLS, with their test RMSE values of 4.12 and 4.16 respectively (ridge and elastic net regression have the same predictive performance). Random forest surprisingly resulted in the highest test RMSE (4.24) among the methods, only slightly better than the intercept-only model (4.268). The training RMSE in the random forest model (1.884) is also much lower than its RMSE, suggestive of overfitting.

Table 4. Comparison of six different methods and their training and test RMSEs

Method	Training RMSE	Test RMSE
Ordinary Lease Squares	3.83463188088652	4.095324290528860
Ridge	4.20170262150334	4.124708372500970
Lasso	4.210207588881210	4.1613810780480500
Elastic Net Regression	4.20170262150334	4.124708372500970
Random Forest	2.0102886157069500	4.247083908858710
Boosted Model	3.5948293315071600	4.001890044125290

Regardless of the different approaches in prediction and test RMSE, these methods all indicate sleep as an important predictor of mental health/depression screening scores. In particular, how

often one feels overly sleepy during the day and/or whether the individual has ever told a doctor about having trouble with sleep are most reliably predictive of mental well-being among the set of selected features in this data. Ratio of family income to poverty also appears in all six methods. Other variables were also shared between different methods, such as gender (OLS, ridge, elnet, boosting), vigorous-intensity recreational activities (OLS, ridge, elnet, random forest, boosting), and self-perception about one's weight (OLS, ridge, elnet, boosting).

Takeaways

Results from the NHANES survey 2017-2020 show that sleep is an important predictor of mental health and depression. The boosted model with the highest predictive performance indicates several items related to sleep, such as the frequency of feeling overly sleepy during the day, whether an individual talked to a health professional about sleep difficulty and hours of sleep during weekdays and on the weekend. Ratio of family income to poverty is also one of the highlighted features in this model. The fact that sleep and income also appear in all other methods (except for hours of weekday and weekend sleep) reinforces the significance of these two factors in mental health. This finding is consistent with the literature on the association between mental disorders, sleep and individual socioeconomic status (SES).

The traditional view that attributes sleep disruptions to be merely a symptom of mental health disorders has been called into question. Instead, research has backed the bidirectional relationship between sleep and mental health (Scott et al., 2017). It is clear that sleep is both the cause and consequence of mental health issues. According to the CDC, more than one third of American adults are sleep-deprived on a regular basis[9]. The Sleep in America Poll for 2020[10] by the National Sleep Foundation found that nearly half of Americans report feeling drowsy between three and seven days per week. Forty percent of adults said that their excessive sleepiness interferes with daily activities at least occasionally (Ibid.). We live in a culture that normalizes long working hours and stigmatizes sleep. However, the accumulated repercussions from inadequate sleep are considerable. Deteriorating physical and mental wellbeing as a consequence of poor sleep can significantly reduce productivity and quality of life.

It is thus time to move from policy conversations to policy actions on changing norms about sleep in the workplace and academic settings. Funding for campaigns and social media have been targeted at educating the public about the importance of sleep. Nevertheless, behavioral change at the individual level will not likely succeed without structural shifts in place so that the blame will be deflected away from the individual. In addition, intake questions about sleep should be intentionally incorporated at medical check-ups and doctor appointments as patients might not always feel comfortable to bring up about problem with their sleep (Grandner & Malhotra, 2015). This will help aid in early detection and prevent arising mental and physical problems in due course.

[9] Center for Disease Control and Prevention (2016). https://www.cdc.gov/media/releases/2016/p0215-enough-sleep.html
[10] The Sleep in America Poll 2020. https://www.sleepfoundation.org/professionals/sleep-america-polls/2020-sleepiness-and-low-action

The findings also further strengthen evidence for the link between SES and mental health. Efforts at reducing inequality have been focused on economic stimulus, job creation and employment training. However, if it's true that SES individuals are at higher risk of mental health issues, additional programming and services specifically on mental well-being are necessary to break the vicious cycle between poor mental wellness, reduced cognitive resources, impaired productivity and subsequently poverty. Initiatives to educate about mental health first-aid and helplines continue to be pivotal. In addition, funding for community resources and activities on mental health—particularly in low-income neighborhoods—will be instrumental in mobilizing community awareness and concerted efforts to improve not only mental well-being but also other interrelated life, social and health outcomes.

Last but not least, the finding on the potential link between gender, vigorous-intensity recreational activities, self-perception about weight and mental health risks can also serve as additional considerations for interventions and programming to push more for support for women, conversations about body image and community initiatives to encourage active physical and recreational activities. Together, the findings corroborate the association between mental health risks and factors at both individual and collective levels. Many structural barriers such as school and workplace norms underlie and influence individual perceptions and behavior related to sleep, physical activity and body self-image. These results thus hope to inform catalytic policy actions to move the needle forward beyond conversations. The time is critical for concrete action to tackle the sleep deprivation epidemic as well as other structural impediments to mental well-being and flourishing.

Limitations

Dataset limitation

Due to COVID-19, data collection for the NHANES 2019-2020 cycle was only partially completed and not nationally representative. As of March 2020, data collection was completed in 18 of 30 locations or primary sampling units and was canceled for the remaining 12 PSUs. The partial 2019-2020 data (2019-March 2020) were combined with the full data set from the previous cycle (2017-2018) to form the current dataset[11].

With a high prevalence of missing data, not all 30 fields in the survey questionnaire could be included. After the removal of missing ("NA") answers, the full dataset would not have met the requirements for analysis. Thus, only a subset of fields (ten) were chosen to form the final dataset. To minimize the instant of cherry-picking, I drew on the literature to form hypotheses before selecting the features.

In addition, the questionnaire and data entry method can be improved. Some questions could be more effective at measuring the phenomenon of interest. For example, instead of asking participants directly about their trouble with sleep, the inquiry was about whether participants

[11] An overview of NHANES survey.
https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/overviewbrief.aspx?Cycle=2017-2020

ever told a doctor about their sleep difficulty. Certain individuals who have sleep problems but never shared such information with medical personnel might have said "No" to this question. The items about physical activity also vaguely asked participants about vigorous moderate-intensity activity at work and at leisure without directly probing whether and what type of physical exercise and activities participants engage in. Mental health and depression screening questions seem to adapt too much clinical criteria that some questions can be too extreme to appropriately measure mental well-being of the non-clinical population (e.g., Moving or speaking slowly or too fast over the past 2 weeks; Thoughts you would be better off dead in the last 2 weeks). The numeric values used to record "Don't know" and "Refused" answers were inconsistent, which posed challenges to data cleaning. In the future, only one number for each category should be utilized uniformly to ensure data quality.

In addition, a number of continuous variables were assigned an upper limit. Variable age was capped at 80 even for participants older than 80 years old. Ratio of family income to poverty used the maximum value of 5 to capture greater values than 5. A value of 14 was used to refer to both 14 or 14+ hours of sleep on the weekday and weekend. Finally, number of hours worked last week in total had an upper limit of 80 hours. The number of hours worked last week might also not necessarily be reflective ò the normal average working hours. The limitation in the upper bound of values, to a certain extent, can interfere with statistical results to detect nuances and variations in the relationship between the feature and response variables.

Finally, given the large number of data fields, it is worth acknowledging the existence of potential confounding variables which can mask or misrepresent the relationship between the response and feature variables. Particularly, in light of the nature of variable selection in LASSO, elastic net regression and tree-based methods, selected features might simply be correlates of other similar variables without necessarily being the most important predictors.

Analysis Limitations

Due to issues with missing data, not all data fields were included in data analysis to ensure the requirements of at least 500 observations. Ten out of 30 fields were selected based on a literature scan to be included in the model. Within each field, I made the deliberate effort to include all items to avoid cherry-picking. However, that is not possible in all cases due to data quality and issues with the questionnaire. Participants under 20 years old were also excluded from analysis as some demographic data was only available for people aged 20 and above. All missing ("NA") responses were omitted from the final dataset.

There was also limitation to the OLS analysis. Due to the categorical nature of many features in the dataset, the statistically significant signal in the output of the OLS regression on all 41 features can only be interpreted as whether a certain level within a variable is significantly different from its default level. In other words, presence of statistical significance does not guarantee to render the selected variables as the most important predictors of mental health risks. More in-depth analysis using F test and multiple linear regression could be explored to

gain an overall assessment of categorical features and further examine potential confounding variables and their interactions.

Despite the utilization of various methods to increase reliable results, analysis results might change with the inclusion of other variables. For example, nutrition has been linked with mental health (Owen & Corfe, 2017; Firth et al., 2020; Zainuddin & Thuret, 2012). However, the entire dietary data was missing for the NHANES 2017-2020 survey. There are also factors specific to COVID-19 that significantly influence mental well-being such as time spent outdoors, social interactions, isolation, anxiety related to loss of employment, increasing racism, issues with safety and future uncertainty. Such information was also not available for analysis.

Last but not least, though different parameters were tuned through various experimentation using cross-validation, the test and training RMSEs need room for improvement. Inherent randomness from splitting data and setting seeds can also influence statistical results. Thus, the audience should interpret the findings with these caveats in mind.

Recommended Follow-ups

To address the aforementioned limitations, additional analyses can be done on 2022 data and beyond. Given the constraints posed by COVID-19, researchers might want to explore other data collection methods such as phone interviews and other online methods. Data collection on nutrition and factors specific to COVID-19 will be helpful for future analysis to gain a fuller picture of predictors of mental health. Missing data will always be an inevitable problem in research. Future analyses could explore imputation methods to deal with missing values[12]. In addition, analyses could be extended to explore predictors of mental health among children and adolescents provided that sufficient data will be available. The survey also misses details on the social life of individuals, such as quality of relationships, social support, social interaction, time spent on social media. Social connection has been shown to predict mental health and well-being and thus should continue to be an essential avenue for research[13] (Klussman et al., 2020). Follow-up analyses could also consider exploring predictors of mental health globally to compare and strengthen scientific findings as well as better inform and tailor policies to each country and different populations around the world.

[12] Towards Data Science. Different Imputation Methods to Handle Missing Data. https://towardsdatascience.com/different-imputation-methods-to-handle-missing-data-8dd5bce97583
[13] NatCen. Predicting wellbeing (2013). https://natcen.ac.uk/media/205352/predictors-of-wellbeing.pdf

Appendix

Explanatory Variables

Below are the 41 explanatory variables used for analysis. Words written in parentheses represent variable names used in R analysis. Unless noted otherwise, all variables are categorical.

Demographic characteristics

- Gender (female): 1 = Female, 2 = Male
- Age (age) - continuous: 0 to 79 = range of values; 80 = 80 years old and older
- Race/Ethnicity (race): Mexican American = 1; Other Hispanic = 2; Non-Hispanic White = 3; Non-Hispanic Black = 4; Non-Hispanic Asian = 6; Other Race - Including Multi-Racial = 7
- Country of birth (born_us): Born in 50 US states or Washington DC = 1; Others = 2
- Education level 20+ (edu): Less than 9th grade = 1; 9-11th grade (Includes 12th grade with no diploma) = 2; High school graduate/GED or equivalent = 3; Some college or AA degree = 4; College graduate or above = 5
- Marital status (marriage_status): Married/Living with Partner = 1; Widowed/Divorced/Separated = 2; Never married = 3
- Ratio of family income to poverty (ratio_income) – continuous: 0-4.98 = Range of values; 5 = Value greater than or equal to 5

Mental health/depression screening (mental_score) – continuous (response variable)

For each of the 10 items below: Not at all = 0; Several days = 1; More than half the days = 2; Nearly every day = 3.

- Have little interest in doing things
- Feeling down, depressed or hopeless
- Trouble sleeping or sleeping too much
- Feeling tired or having little energy
- Poor appetite or overeating
- Feeling bad about yourself
- Trouble concentrating on things
- Moving or speaking slowly or too fast
- Thoughts you would be better off dead
- Difficulty these problems have caused

Smoking

- Age started smoking cigarettes regularly (age_first_smoke) – continuous
- Do you now smoke cigarettes? (now _smoke): Every day = 1; Some days = 2; Not at all = 3

Physical activity

- Minutes sedentary activity (min_sedentary) – continuous
- Vigorous work activity (vigor_work):
  Next I am going to ask you about the time {you spend/SP spends} doing different types of physical activity in a typical week. Think first about the time {you spend/he spends/she spends} doing work. Think of work as the things that {you have/he has/she has} to do such as paid or unpaid work, household chores, and yard work. Does {your/SP's} work involve vigorous-intensity activity that causes large increases in breathing or heart rate like carrying or lifting heavy loads, digging or construction work for at least 10 minutes continuously? Yes = 1; No = 2
- Moderate work (moderate_work):
  Does {your/SP's} work involve moderate-intensity activity that causes small increases in breathing or heart rate such as brisk walking or carrying light loads for at least 10 minutes continuously? Yes = 1; No = 2
- Walk or bike (walk_bike):
  The next questions exclude the physical activity at work that you have already mentioned. Now I would like to ask you about the usual way {you travel/SP travels} to and from places. For example to school, for shopping, to work. In a typical week {do you/does SP} walk or use a bicycle for at least 10 minutes continuously to get to and from places? Yes = 1; No = 2
- Vigorous recreational activities (vigor_rec):
  The next questions exclude the work and transport activities that you have already mentioned. Now I would like to ask you about sports, fitness and recreational activities. In a typical week {do you/does SP} do any vigorous-intensity sports, fitness, or recreational activities that cause large increases in breathing or heart rate like running or basketball for at least 10 minutes continuously? Yes = 1; No = 2
- Moderate recreational activities (moderate_rec):
  In a typical week {do you/does SP} do any moderate-intensity sports, fitness, or recreational activities that cause a small increase in breathing or heart rate such as brisk walking, bicycling, swimming, or volleyball for at least 10 minutes continuously? Yes = 1; No = 2

Sleep

- Sleep hours – weekdays or workdays (sleep_weekday) – continuous
  2 = Less than 3 hours; 3 – 13.5 = range of values; 14 = 14 hours or more
- Sleep hours – weekends or non-workdays (sleep_weekend) – continuous
  2 = Less than 3 hours; 3 – 13.5 = range of values; 14 = 14 hours or more
- Ever told a doctor about trouble sleeping (dr_sleep): 1 = Yes; 2 = No
- How often feel overly sleepy during day? (sleepy)
  In the past month, how often did {you/SP} feel excessively or overly sleepy during the day?): Never = 0; Rarely - 1 time a month = 1; Sometimes - 2-4 times a month = 2; Often- 5-15 times a month = 3; Almost always - 16-30 times a month = 4

Medical history

Has a doctor ever told you that you have…(Yes = 1; No = 2)

- Asthma (told_asthma)
- Arthritis (told_arthritis)
- Coronary heart disease (told_coronary)
- Liver condition (told_liver)
- COPD, emphysema or bronchitis (told_copd)
- Stroke (told_stroke)
- Thyroid (told_thyroid)
- Cancer (told_cancer)

During the past 12 months {have you/has s/he} ever been told by a doctor or health professional to (Yes = 1, No = 2)…

- control {your/his/her} weight or lose weight? (told_weight)
- increase {your/his/her} physical activity or exercise? (told_exe)
- watch or reduce the amount of sodium or salt in {your/his/her} diet? (told_salt)
- watch or reduce the amount of fat or calories in {your/his/her} diet? (told_fat)

Are you now (Yes = 1; No = 2)…

- Controlling or losing weight? (now_lose_weight)
- Increasing exercise? (now_increase_ex)
- Reducing salt in diet? (now_reduce_salt)
- Reducing fat in diet? (now_reduce_fat)

Diabetes

- Doctor told you have diabetes (dr_diabetes): Yes = 1; No = 2

Blood pressure and cholesterol

- Ever told you had high blood pressure (dr_highbp): Yes = 1; No = 2
- Doctor told you - high cholesterol level (dr_highchol): Yes = 1; No = 2

Occupation

- Hours worked last week in total all jobs (hours_worked) – continuous
  5 = 1-5 hours; 6-78 = range of values; 80 = 80 hours or more

Weight history

- How do you consider your weight? (weight_self_percept): Overweight = 1; Underweight = 2; About the right weight = 3
- Like to weigh more, less or same (want_lose_weight): More = 1; Less = 2; Stay the same = 3

References

Alroomi, A. S., & Mohamed, S. (2021, April). Predictors of mental health and fatigue among isolated oil and gas workers. In *Safety and Reliability* (Vol. 40, No. 2, pp. 80-98). Taylor & Francis

Axon, D. R., & Chien, J. (2021). Predictors of Mental Health Status among Older United States Adults with Pain. *Behavioral Sciences*, *11*(2), 23

Corrigan PW, Morris SB, Michaels PJ, Rafacz JD, Rüsch N. Challenging the Public Stigma of Mental Illness: A Meta-Analysis of Outcome Studies. Psychiatric services (Washington, DC) 2012;63(10):963–73. 10.1176/appi.ps.201100529. [PubMed] [CrossRef] [Google Scholar]

Duffy, A., Keown-Stoneman, C., Goodday, S., Horrocks, J., Lowe, M., King, N., Saunders, K. (2020). Predictors of mental health and academic outcomes in first-year university students: Identifying prevention and early-intervention targets. *BJPsych Open, 6*(3), E46. doi:10.1192/bjo.2020.24

Farajzadeh, A., Dehghanizadeh, M., Maroufizadeh, S., Amini, M., & Shamili, A. (2021). Predictors of mental health among parents of children with cerebral palsy during the COVID-19 pandemic in Iran: A web-based cross-sectional study. *Research in Developmental Disabilities*, *112*, 103890).

Firth, J., Gangwisch, J. E., Borisini, A., Wootton, R. E., & Mayer, E. A. (2020). Food and mood: How do diet and nutrition affect mental wellbeing? *BMJ*, *369*

Grandner, M. A., & Malhotra, A. (2015). Sleep as a vital sign: why medical practitioners need to routinely ask their patients about sleep. *Sleep health*, *1*(1), 11–12. https://doi.org/10.1016/j.sleh.2014.12.011

Hennein, R., Mew, E. J., & Lowe, S. R. (2021). Socio-ecological predictors of mental health outcomes among healthcare workers during the COVID-19 pandemic in the United States. *PloS one*, *16*(2), e0246602

Hjorth CF, Bilgrav L, Frandsen LS, Overgaard C, Torp-Pedersen C, Nielsen B et al. Mental health and school dropout across educational levels and genders: a 4.8-year follow-up study. BMC Public Health 2016;16:976 10.1186/s12889-016-3622-8. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

Klussman, K., Nichols, A. L., Langer, J., & Curtin, N. (2020). Connection and disconnection as predictors of mental health and wellbeing. *International Journal of Wellbeing*, *10* (2).

Ljungqvist I, Topor A, Forssell H, Svensson I, Davidson L. Money and mental illness: A study of the relationship between poverty and serious psychological problems. *Community Mental Health J* 2016; 52(7), 842–50. [PubMed] [Google Scholar]

Nochaiwong, S., Ruengorn, C., Thavorn, K., Hutton, B., Awiphan, R., Phosuya, C., Ruanta, Y., Wongpakaran, N., & Wongpakaran, T. (2021). Global prevalence of mental health issues among the general population during the coronavirus disease-2019 pandemic: A systematic review and meta-analysis. *Scientific Reports*, *11*(1), 10173. https://doi.org/10.1038/s41598-021-89700-8

Owen, L., & Corfe, B. (2017). The role of diet and nutrition on mental health and wellbeing. *Proceedings of the Nutrition Society*, *76*(4), 425-426

Santini, Z. I., Stougaard, S., Koyanagi, A., Ersbøll, A. K., Nielsen, L., Hinrichsen, C., ... & Koushede, V. (2020). Predictors of high and low mental well-being and common mental disorders: Findings from a Danish population-based study. *European Journal of Public Health*, *30* (3), 503-509.

Scott, A. J., Webb, T. L., & Rowse, G. (2017). Does improving sleep lead to better mental health? A protocol for a meta-analytic review of randomised controlled trials. *BMJ open*, *7*(9), e016873.

Trabelsi, K., Ammar, A., Masmoudi, L., Boukhris, O., Chtourou, H., Bouaziz, B,... (2021). Sleep quality and physical activity as predictors of mental wellbeing variance in older adults during COVID-19 lockdown. *International Journal of Environmental Research and Public health*, *18*(8), 4329.

Wu, T., Jia, X., Shi, H., Niu, J., Yin, X., Xie, J., & Wang, X. (2021). Prevalence of mental health problems during the COVID-19 pandemic: A systematic review and meta-analysis. *Journal of Affective Disorders*, *281*, 91-98.

Zainuddin, M. S. A., & Thuret, S. (2012). Nutrition, adult hippocampal neurogenesis and mental health. *British Medical Bulletin*, *103*(1), 89-114.)