Psychological Test and Assessment Modeling, Volume 61, 2019 (4), 471-493

Predicting post-experiment fatigue among healthy young adults: Random forest regression analysis

Eun-Young Mun[1] *& Feng Geng*[2]

Abstract

The current study utilized a random forest regression analysis to predict post-experiment fatigue in a sample of 212 healthy participants (mean age = 20.5, *SD* = 2.21; 52% women) between the ages of 18 and 30 following a mildly stressful experiment. We used a total of 30 features of demographic variables, lifestyle variables, alcohol and other drug use behaviors and problems, state anxiety and depressive symptoms, and physiological indicators that were lab assessed or self-reported. A random forest regression analysis with 10-fold cross-validation resulted in accurate prediction of post-experiment fatigue (R^2 equivalent = 0.93) with the average "out-of-bag" (OOB) R^2 = 0.52. Not surprisingly, self-reported pre-experiment fatigue was the most important variable (54%) in the prediction of post-experiment fatigue. Feeling anxious (state anxiety) pre- and post-experiment (3%, 7%), feeling less vigorous post experiment (3%), systolic and diastolic blood pressure (3%, 2%) and LF HRV (2%) assessed at baseline, and self-reported alcohol-related problems (3%) and sleep (2%) additionally contributed to the prediction of post-experiment fatigue. Other remaining input variables had relatively minimal importance. Substantively, this study suggests that complex interactions across multiple systems domains that support regulation may be linked to fatigue. A random forest regression analysis can relatively easily be implemented with a built-in cross-validation function and reveal a web of connections undergirding health behavior and risks.

Keywords: fatigue, stress, regulation, random forests, machine learning

[1] *Correspondence concerning this article should be addressed to:* Eun-Young Mun, PhD, Department of Health Behavior and Health Systems, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., EAD 709, Fort Worth, TX 76107-2699, USA; email: eun-young.mun@unthsc.edu.

[2] Northwestern University, Evanston, USA

Introduction

Fatigue is an important health indicator variable, but has not been studied frequently perhaps due to its generality or lack of disease specificity (Tiesinga, Dassen, Halfens, & van den Heuvel, 1999). Chronic fatigue has been more commonly studied for patients with neuromuscular diseases or more generally patients who underwent medical interventions. However, fatigue can be situational and dynamic, as well as chronic. Fatigue as a momentary measure may reflect one's physical and mental state in response to surrounding environmental demands, and may be used as an indicator of psychological stress response (e.g., Travis et al., 2018). In a recent neuroimaging study, self-reported momentary fatigue assessed using the Profile of Mood States Brief Form (POMS; McNair, Lorr, Heuchert, & Droppleman, 1989), but not other measures of stress or chronic fatigue, was significantly correlated with lower gray matter brain volume among a small sample of healthy adults (Kokubun et al., 2018). Therefore, there is a need to better understand fatigue as a state-like health outcome variable to assess one's ability to handle discomfort in response to external and internal stimulations.

For ostensibly healthy young adults, being able to withstand mild physical discomfort and psychological distress that stems from participating in an experimental study may signal one's adaptability and better regulation of stress and mood. In this study, we focus on fatigue reported as a momentary mood state. We examined demographic and lifestyle variables such as age, sex, body mass index, physical activity, sleep, and alcohol and drug use behavior, as well as physiological measures, such as mean heart rate (HR), indices of heart rate variability (HRV) and cortisol levels observed in the lab as part of a larger study in a sample of healthy young adults. The experiment, which lasted for approximately 1.5 hours, was aimed at examining how individuals respond to various external stimuli that are pleasant or stressful using a multi-modal assessment approach, including voluntary self-report measures, and involuntary physiological and hormonal indices. The physiological and hormonal indices tap into several self-regulatory systems in the brain through peripheral systems that figure prominently in the regulation of emotion and stress (Gunnar & Davis, 2003), including the autonomic nervous system (ANS) and the limbic-hypothalamic-pituitary-adrenocortical (L-HPA) system. The nature of physical discomfort and psychological distress involved sitting in a comfortable chair for up to 45 minutes without movements for four electrocardiogram (ECG) recording sessions (two separate five min sessions, and two consecutive five min sessions) for a total 20 minutes. They were allowed to move their arms and legs during a short break between sessions but were generally limited in their movement due to being connected to sensors. During ECG recording, they were asked to remain still. In addition, participants were subjected to a topical application of common consumer pilot products on their skin (e.g., 1.5% menthol in propylene glycol that are frequently found in skin and hair products), and self-reported potentially sensitive information, such as alcohol and drug use, on a desktop computer.

To leverage emerging computational capabilities, we used a machine learning approach. There has been a growing interest in big data and data science in the field of psychology and clinical research. For example, while introducing a recent special issue in the journal

Psychological Methods, its editors declared that “big data or data science is here to stay, with or without psychology” (Harlow & Oswald, 2016, p. 2). Improving big data capabilities for biomedical and clinical research is also one of the major strategic plans of the National Institutes of Health for 2016-2020 (National Institutes of Health, 2018). Unfortunately, data applications using these techniques are scarce in the field of psychology. In the current study, we utilized a random forest regression and classification algorithm, an intensive data-driven machine learning algorithm engineered to detect complex relationships among variables. Given the scarcity of data examples, we used the quoted terms in the current study to highlight overlap in terminology.

Random forests (Breiman, 2001) is a nonparametric, ensemble learning algorithm used for regression for a numerical “target” (or dependent or outcome variable) and classification for a binary outcome variable using “features” (or independent or input variables), which has been shown to yield significant improvements in accuracy in prediction and classification, compared to other algorithms. Random forests regression and classification shares the same recursive partitioning method with classification and regression trees (CART; “grow” or “build” the tree, but without pruning) that is designed to maximize within-node homogeneity at each step of growing the tree (or “partitioning or “splitting” each node). CART is relatively simple in its implementation and interpretation because it is based on a single tree that is pruned back based on pre-specified stopping rules. However, CART can be unstable with its single tree variance larger than the variance of random forests (Dankowski, & Ziegler, 2016; Steingrimsson, Diao, & Strawderman, 2019). Random forest regression reduces the variance of the random forest and achieves prediction accuracy via random “feature selection” at each split and “bagging” of a large number of different trees in a forest (i.e., bootstrap aggregating). In other words, a random forest is random in two ways – a random set of “features” and a random set of “training data” (Grömping, 2009), which reduces the correlation between the individual trees and consequently reduces the variance of the random forest (Breiman, 2001). Random forests have been increasingly popular because of their high accuracy, suitability for data sets with high *p* and low *n* problems, ability to handle highly correlated data and complex interactions (Archer & Kimes, 2008; Grömping, 2009; Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008).

Random forests algorithm uses the “bootstrap” method (Efron, 1979), a method that has been extensively used to assess the uncertainty and accuracy of scientific results across many scientific fields, with its developer Bradley Efron recognized with the “International Prize in Statistics” award in November 2018 (American Statistical Association, 2018). Bootstrapping is a resampling method that draws samples (called “bootstrap samples”) from an estimate of the population (i.e., “a sample”) instead of the population to learn about the sampling distribution. This is because we rarely have multiple samples from the population. Rather, we have only one sample. Therefore, to learn about the population and its sampling distribution, we use the “bootstrap distribution” (see Cochran, 2018 for a short accessible review). Bootstrapping proceeds by taking repeated samples with replacement from a sample, followed by computing the statistic of interest (e.g., mean or regression coefficient) from each bootstrap sample. Then, the resulting statistics are collected and their distribution is called the bootstrap distribution. The bootstrap distribution can be useful for obtaining information about the sampling distribution of the population, such as the spread and shape, but not for estimating the cumulative distribution function or quantiles of a sample statistic because the bootstrap distribution can be biased and also because it is narrow by a factor of $\sqrt{(n-1)/n}$ for the sample mean and other statistics (Hesterberg, 2015). In the field of psychology, bootstrapping has been frequently utilized for obtaining confidence intervals for indirect mediation effects. However, the bootstrap methods (both bias-corrected and bias-corrected and accelerated) have been shown to suffer from inflated Type I and Type II error rates when testing mediation effects, especially in small samples (N < 100; Koopman, Howe, Hollenbeck, & Sin, 2015). In the context of random forests, bootstrapped samples are used to develop nonparametric regression or classificaiton models, which are subsequently aggregated in terms of prediction and prediction errors of the aggregated model.

A random forest regression proceeds as follows. First, we draw bootstrap samples (e.g., 2/3) for a number of trees (e.g., 100 trees in a forest) from original "training" data. For each bootstrap sample, a portion of observations is left out as an "out-of-bag" sample (e.g., 1/3). Second, using each training data set, we grow a regression tree and draw a random subset of "features" (predictors) that explain the outcome variable. Third, we use "out-of-bag" observations that were left out of the bag to evaluate the subset of trees (models) in the forest that did not include the "out-of-bag" observations to generate "OOB" prediction estimates (OOB R^2 for "out-of-bag" observations). The predicted "out-of-bag" observations are averaged to obtain OOB R^2 and OOB mean squared error (MSE) score. Thus, the accuracy of a random forest's prediction can be estimated from OOB data as:

$$OOB\ MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{\hat{y}}_{i\ OOB})^2$$

where y_i denotes the observed "target" (outcome variable) value for participant i and $\bar{\hat{y}}_{i\ OOB}$ indicates the average prediction for this i th participant from all tress for which this observation has been left out as an "out-of-bag" observation (Grömping, 2009). The OOB MSE quantifies the generalization error – in other words, the error rate of the regression (classifier for a binary outcome) trees obtained from the "training data set" on "out-of-bag" samples. It is because with enough trees, OOB prediction error converges to the fit of a forest on similar new data (e.g., "test data set"). The "importance" of a variable can be examined by looking at increases in prediction error when "out-of-bag" data for that variable is permuted (i.e., null) when all others remain the same (i.e., permutation importance) or by examining the mean reduction in impurity (i.e., Gini importance). There are other importance metrics but these two are most commonly utilized. The importance of a variable may reflect possible complex interactions with other variables as well as dependency among independent variables in a model (Breiman, 2001; Liaw & Wiener, 2002).

- A random forest regression can proceed for $b = 1, \ldots, B$: Draw a bootstrap sample $\mathbf{X}_b$ from the training data set;
- Grow an individual regression tree f_b to the bootstrapped data $\mathbf{X}_b$ and output the ensemble of trees;
- Predictions for "out-of-bag" observations can be made by averaging the predictions from all individual trees on the "out-of-bag" observations x : $\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$.

Random forest analysis is computationally intense, but better takes into account complex higher-order interactions that exist in data without any need to explicitly specify them (Grömping, 2009), provides a more accurate prediction model, has built-in estimates of error, strength, correlation and variable importance, is more robust to outliers and noise, and is fast (Breiman, 2001).

As is typical in studies utilizing machine learning algorithms, the current investigation was exploratory in nature. We sought to uncover hidden complex patterns in data. Random forests analysis is particularly well suited for this purpose because it is not necessary to reduce the number of variables before analysis and there is no need to specify interactions (Archer & Kimes, 2008; Grömping, 2009). We used sklearn.ensemble algorithm from Scikit-Learn (v0.22) developed for Python (Python Core Team, 2019). Scikit-Learn is a free software machine learning library for the Python programming language (for more information, https://scikit-learn.org/stable/faq.html).

Methods

Participants

A total of 212 apparently healthy participants (mean age = 20.5, *SD* = 2.21; 52% women) between the ages of 18 and 30 without any known skin or respiration problems were tested. The majority of participants were Asian 46%; 34% were Caucasian, 9% were African American, and 11% identified as other or missing (3 cases). Participants were recruited into the study from 2011 to 2012 through word of mouth, university bulletin board fliers, website advertisement, and advertisement in the university newspaper. Those who expressed interest were contacted via phone to complete a brief screening interview in which they were assessed on the exclusion criteria. Individuals with physical conditions such as diabetes, respiratory problems, kidney or liver disease, or cardiovascular problems, or those who were pregnant or planning to become pregnant were excluded from participation. Finally, those who had high blood pressure (over 140 systolic and/or over 90 diastolic) or who were over- or under-weight (20% above or below from ideal weight) were excluded. Eligible participants were then scheduled for an experimental session between 10 am and 4 pm to control for circadian rhythms. At this time, they were instructed to abstain from any alcohol or drugs for 24 hrs prior to testing, and asked to fast for 1 hour prior to testing.

Procedures

Upon arrival, informed consent was obtained, and blood pressure, height, and weight were assessed. Any individuals who did not meet inclusion criteria based on these measures were thanked for their time, and compensated with $10. The remaining participants continued with the experimental session that consisted of both self-report and physiological assessments. They reported all self-report measures on a computer screen. Baseline self-report measures were completed prior to three task sessions. All subjects were tested individually for: (1) a 5-min baseline; (2) a 5-min placebo (not reported), and (3) a 10 min task session, during which two successive 5 min recordings took place. Each task session was approximately 10-15 min apart during which subjects completed a short questionnaire about their experience and mood states, and provided a saliva sample. In all three sessions, participants viewed colored objects (rectangles, circles, squares, or triangles) at the rate of one object per 10 s and silently counted the number of certain colored objects. This standardized, cognitively low-demand task (Jennings, Kamarck, Stewart, Eddy, & Johnson, 1992) is to equate the influence of cognitive load on HRV across participants (Jorna, 1992; Sloan, et al., 1994). A saliva sample (2 ml) was collected following this recording.

To measure post-experiment fatigue, we used the physiological and hormonal indices observed from the 2nd phase of the final 3rd task session, and self-report measures. The entire experiment lasted for approximately 1.5 hours, and participants were paid $20-$25 for their time (subject payment increased to $25 midway through the study to boost participation). All study procedures were approved by a university Institutional Review Board (Protocol 11-278M).

Data processing

The ECG record and respiration frequency were collected at a rate of 2,000 samples per s by a Powerlab acquisition system (ADInstruments, Colorado Springs, CO). LabChart 7.2 (ADInstruments, Colorado Springs, CO) was used for analyses and calculation of physiological indices. For HRV analysis, beat-to-beat RR intervals (RRI) in heart rhythms were recorded, edited, and segmented into 5 min blocks for fast Fourier transformation spectral analysis.

Saliva samples were frozen at -20°C until processed for analysis. Samples were then thawed and initially centrifuged at 3500 x g for 30 min to precipitate the insoluble mucins. To assess cortisol levels, 25 µl aliquots of clarified saliva were assayed in duplicate with Enzyme Immunoassay (EIA) reagents from Salimetrics, State College, PA (High Sensitivity Salivary Cortisol EIA kit). According to manufacturer, intra-assay and inter-assay variations are 3.4 and 3.7% respectively. Lower limit of sensitivity is < 0.003 µg/dl. Plates were read at 450 nm, corrected by reading at 630 nm on a Bio-Tek microplate reader. Approximately 90% of the saliva samples were analyzed for cortisol.

Measures

Table 1 reports descriptive statistics of all variables. Figure 1 provides bivariate correlations among all pairs of the 31 variables under investigation.

Table 1:
Descriptive Statistics of All 31 Variables

Variable	*Mean*	*SD*
Baseline individual characteristics		
Men (1; women = 0 [SEX])	0.48	0.50
Age in years (Age)	20.54	2.21
Hours of sleep during the past night (Sleep)	7.09	1.64
Light to moderate leisure physical activities (FRLIX)	5.53	1.66
Body mass index (BMI)	22.54	3.86
Cigarette smoking frequency in the past month (CIGA)	0.84	1.58
Marijuana use frequency in the past month (MARI)	0.97	1.56
Number of drinks in a typical week in the past month (DRKS)	7.04	11.96
Heavy episodic drinking frequency in the past month (BINGE)	1.15	2.23
Alcohol-related problems (RAPI)	3.16	5.84
Depressive symptoms (CESD)	12.70	8.15
Systolic blood pressure at baseline (BPS)	121.29	12.44
Diastolic blood pressure at baseline (BPD)	75.42	7.88
Lab assessed psychophysiological measures		
Mean heart rate at baseline (HR10)	72.84	10.77
Mean heart rate during the last session (HR31)	72.08	9.66
HRV - SDNN at baseline (SDNN10)	59.73	24.52
HRV - SDNN during the last session (SDNN31)	67.53	26.79
HRV - LF/HF ratio at baseline (LFHF10)	1.59	2.12
HRV - LF/HF ratio during the last session (LFHF31)	2.35	2.47
HRV - LFHRV at baseline (LFHRV10)	6.68	0.87
HRV - LFHRV during the last session (LFHRV31)	7.14	0.86
HRV - HFHRV at baseline (HFHRV10)	6.63	1.13
HRV - HFHRV during the last session (HRHRV31)	6.68	1.10
Cortisol at baseline (Corti10)	0.20	0.14
Cortisol after the last session (Corti31)	0.17	0.11
Self-reported state anxiety and mood		
Pre-experiment STAI state anxiety (ASTAI)	36.83	9.14
Post-experiment STAI state anxiety (CSTAI)	38.33	9.91
Pre-experiment POMS vigor (apomsVIG)	4.45	3.90
Post-experiment POMS vigor (cpomsVIG)	2.96	3.87
Pre-experiment POMS fatigue (apomsFAT)	4.21	4.08
Post-experiment POMS fatigue (cpomsFAT)	4.91	4.97

Note. Variable names are in parentheses.

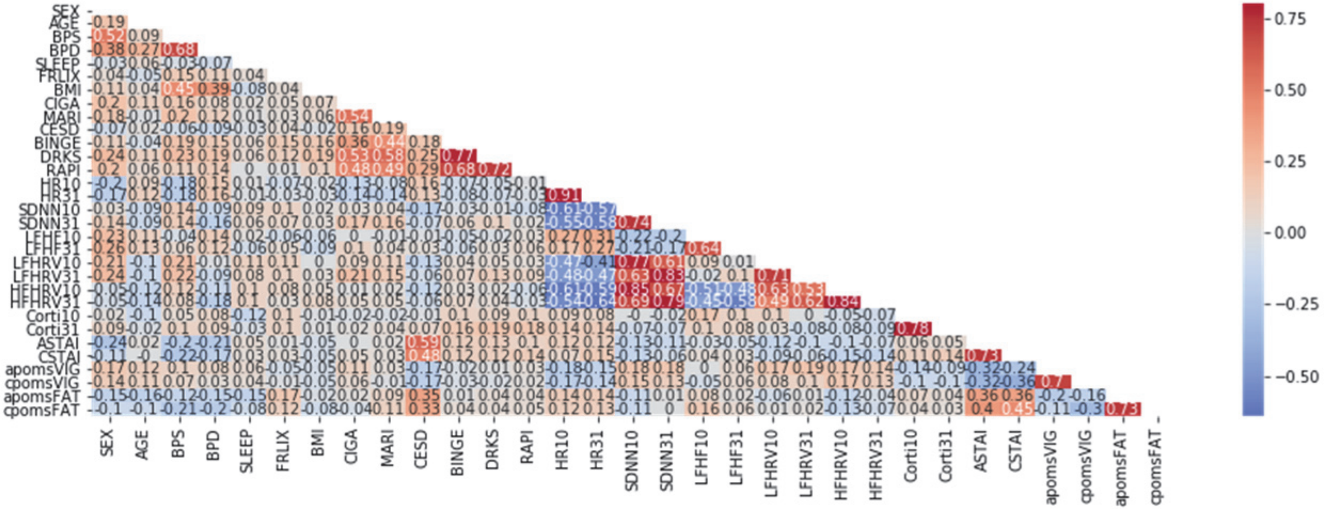

Figure 1:
Bivariate correlations of all pairs of variables (features). The target variable, self-reported post-experiment fatigue was most strongly correlated with the following variables in the order of magnitude: Pre-experiment POMS fatigue (0.73), post-experiment STAI state anxiety (.45), pre-experiment STAI state anxiety (0.40), CESD depressive symptoms (0.33), post-experiment POMS vigor (-.30), systolic blood pressure at baseline (-.21), diastolic blood pressure at baseline (-.20), LF/HF ratio (.16) at baseline, mean heart rate during the last task session (.13), HF HRV at baseline (-.13), mean heart rate at baseline (0.12), and light to moderate physical activity (0.12)

Mood. The Profile of Mood States Brief Form (POMS; McNair, Lorr, Heuchert, & Droppleman, 1989) was used to assess mood. This measure includes a list of 30 mood states (e.g., tense, angry, worn out) across six subscales including tension-anxiety, anger-hostility, fatigue-inertia, depression-dejection, confusion-bewilderment, and vigor-activity. Participants were presented with the list and asked to indicate their mood at that moment on a 5-point scale ranging from 0 = *Not at all* to 4 = *Extremely*. Items within each subscale were summed to create a score for that respective mood. We report two subscale scores Fatigue (inertia) and Vigor (activity).

Anxiety. State anxiety was assessed using the State-Trait Anxiety Inventory for Adults (STAI; Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983). This measure includes two forms, Y-1 and Y-2, each with 20 items. Form Y-1 assesses feelings of anxiety in the moment, or state anxiety. Participants were asked to indicate their feelings on a 4-point scale ranging from *not at all* (1) to *very much so* (4). Form Y-2 measures general feelings of anxiety, or trait anxiety. Similar to form Y-1, students were asked to read the items and indicate their feelings on a 4-point scale ranging from *almost never* (1) to *almost always* (4). We summed items specific to Y-1 form to create state anxiety scores.

Depressive symptoms. Depressive symptoms were assessed using the Center for Epidemiologic Studies Depression Scale (CES-D; Radloff, 1977). The CES-D consists of 20 self-report items and provides a measure of current depressive symptoms. Participants were asked to indicate how many days during the past week they had experienced the emotions or behaviors indicated in each of the items. The response options ranged from 0 = *rarely or none of the time* to 3 = *most or all of the time*. Items were prorated for missing response and summed to create a total CES-D scale score.

Alcohol consumption, alcohol-related problems, and cigarette and marijuana use. Alcohol use quantity in a typical week in the past month was assessed using the Daily Drinking Questionnaire (DDQ; Collins, Parks, & Marlatt, 1985). Heavy episodic drinking was assessed by asking participants the number of times they consumed five drinks or more (four or more for women) within two hours over the past month. Alcohol-related problems for the past three months were assessed using the 18-item version of the Rutgers Alcohol Problem Index (RAPI; White & Labouvie, 1989, 2000). The RAPI assesses the extent to which participants' daily functions and social relationships were affected by drinking and whether they experienced a higher alcohol tolerance or a blackout. For the 18 items presented, participants were asked to respond how often an outcome had occurred, in the past three months, using the following options: 0 = *None*; 1 = *1-2 times*; 2 = *3-5 times*; 3 = *More than five times*. Cigarette smoking was self-reported using a single question that asked about the frequency of tobacco use during the past month: 0 = *Never used cigarettes*; 1 = *Not used in the past month*; 2 = *Once a month*; 3 = *Two or three times a month*; 4 = *Once or twice a week*; 5 = *Three or four times a week*; 6 = *Every day or nearly every day*. Marijuana use frequency was assessed using a single item question: How often have you used Marijuana or Hashish in the past month? Response options ranged from 0 = *Never used marijuana or hashish*; 1 = *Not used in the past month*; 2 = *Once a month*; 3 = *Two or three times a month*; 4 = *Once or twice a week*; 5 = *Three or four times a week*; 6 = *Every day or nearly every day*.

Lifestyle variables. Questions about physical activity, sleep, and other health-related behaviors were taken from the National Health Interview Survey (United States Department of Health and Human Services, 2009). Sleep was assessed by a single item "How many hours of sleep did you get last night?" Participants answered in an open ended response field, which ranged from 0 to 11.5 hours. Physical activity was assessed by a single item "How often do you do light or moderate leisure-time physical activities for at least 10 minutes that cause only light sweating or a slight moderate increase in breathing or heart rate?" The response option ranged from 0 = *Never*, 1 = *One or two days in the year*, 2 = *Several days in the year*, 3 = *One day a month*, 4 = *Two or three days a month*, 5 = *One day a week*, 6 = *Two or three days a week*, 7 = *Four to six days a week*, and 8 = *Everyday*. Body Mass Index (BMI) was calculated by dividing one's weight in kilograms, by the square of his or her height in meters. Participants' height and weight were assessed in the lab by trained research assistants.

Blood pressure, mean HR, HRV indices, salivary cortisol. Systolic and diastolic blood pressure was assessed in the lab by trained research assistants and recorded. ECG was assessed to evaluate mean heart rate (HR) and heart rate variability (HRV). HRV assesses variability in the beat-to-beat intervals in heart rhythms (i.e., R to R intervals or RRI). The power spectrum of the RRI in the low frequency (0.05-0.15 Hz) range (LF HRV index) reflects the baroreflex condition that is implicated in active emotion regulation. High frequency HRV index quantifies the strength of the signals in the high frequency range (0.15 to 0.4 Hz), which is often associated with positive emotions. The ratio of LF to HF assesses the ratio between the sympathetic nervous system to parasympathetic nervous system activity. A high LF/HF ratio is generally interpreted as sympathetic dominance, which occurs in response to acute stress or with parasympathetic withdrawal. Standard deviation of normal-to-normal intervals (SDNN) assesses total variability within each of the 5 min sessions (see Appelhans & Luecken, 2006; Berntson et al., 1997; Shaffer & Ginsberg, 2017 for more information on these indices). Salivary cortisol is a well-known biomarker of acute psychological stress as well as various chronic stress related conditions (Dobler et al., 2019).

Results

There were 1% missing data for 31 variables across 212 participants. For this missing data, we used a regression-based "IterativeImputer" algorithm for Python. We then preprocessed all variables so that they were scaled (see Figure 2). This is akin to standardizing all variables in cluster analysis (see Mun, von Eye, Bates, & Vaschillo, 2008; Mun, Windle, & Schainker, 2008).

We used a 10-fold cross validation for a total of 80 trees per analysis. The selection of 80 trees was based on the analysis that examined the effect of number of tree estimators on prediction error rates. Figure 3 shows that the OOB error rate precipitously decreased as the number of trees increased but reached a plateau around 80. The two lines indicate different performance depending on the choice of maximum features (variables). We

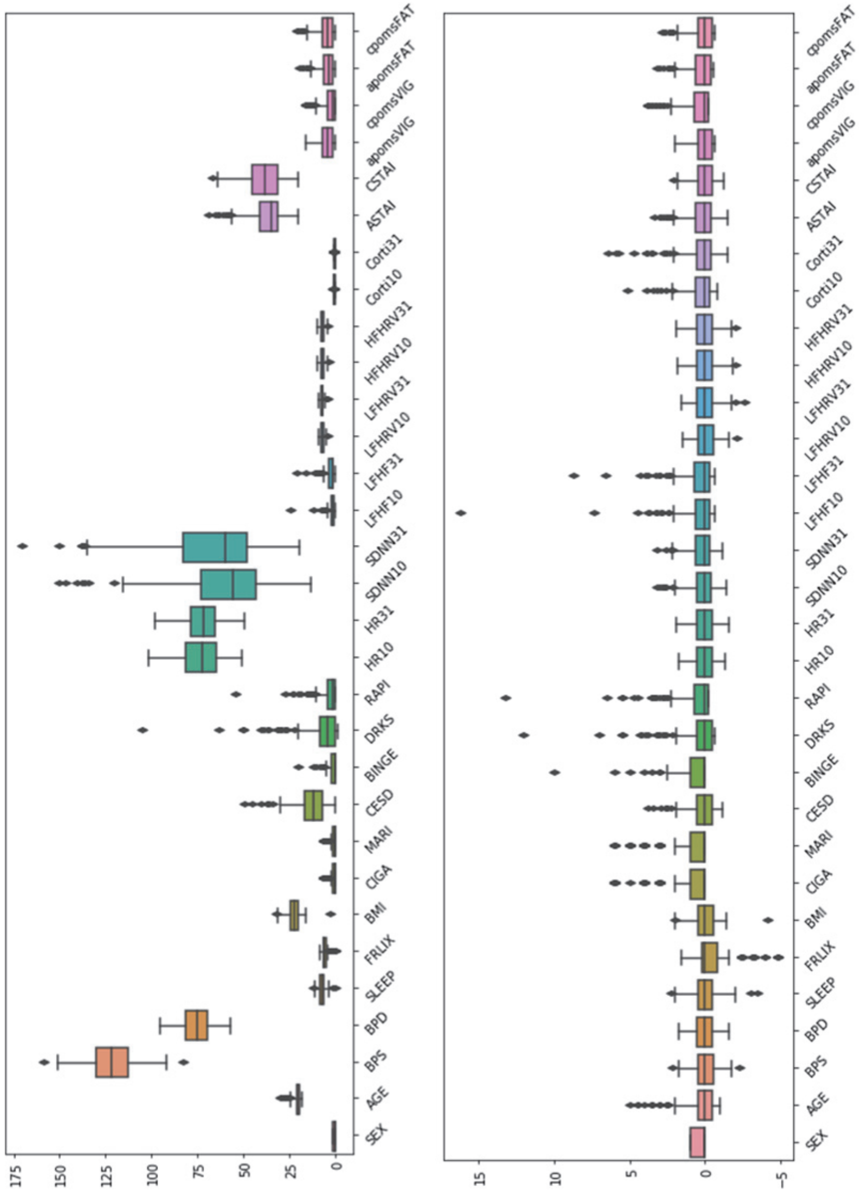

Figure 2: Before (top) and after (bottom) preprocessing and scaling data for analysis

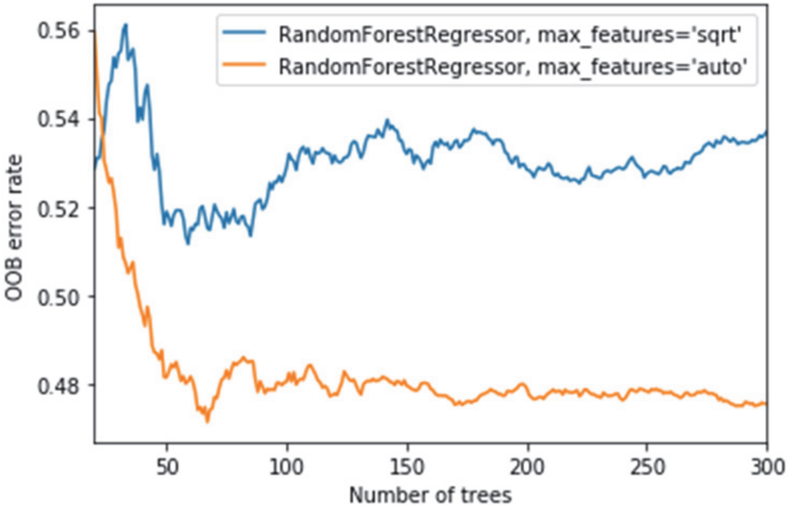

Figure 3:
OOB error rate drops as the number of trees (estimators) increases

used "auto," which means that the model could automatically select up to all 30 features, whereas "sqrt" would select 5 or 6 per splitting (a square root of 30 = 5.5). We also examined the effect of number of estimators based on another error metric (not shown), which showed an almost identical pattern of a rapid decrease in error followed by a small gain in error reduction after about 75 estimators. The proportion of the variance explained (R^2 equivalent) from the "training" data was 0.93. The prediction for "out-of-bag" observations (OOB R^2) was 0.52. The OOB R^2 score quantifies the average prediction for "out-of-bag" observations by using the prediction model from all of the trees that did not include the "out-of-bag" observations. These two numbers 0.93 and 0.52 can be seen as upper and lower bounds of the prediction when the forest is applied to new data (SAS Institute, 2016). Average root mean squared error (RMSE) to evaluate the model was 0.56 (*SD* = .10) across 10 cross validation sets.

We then derived "importance" of variables across the trees (see Figure 4 and also Table 2). This measure is scaled so that the sum of all feature importance scores becomes 1 (or 100%), which suggests the relative importance of a variable or feature among all input variables. Variable importance can be calculated by permuting values of a given variable in each tree's "out-of-bag" sample while keeping the rest the same, which results in increased prediction error and is expressed as a percent increase in mean square error (i.e., permutation importance; Liaw & Wiener, 2002). Scikit-Learn examines the mean decrease in impurity on average across all trees in the forest when a variable was used for the nodes (i.e., Gini importance; Géron, 2019, p. 200; see also Scheidel, 2018).

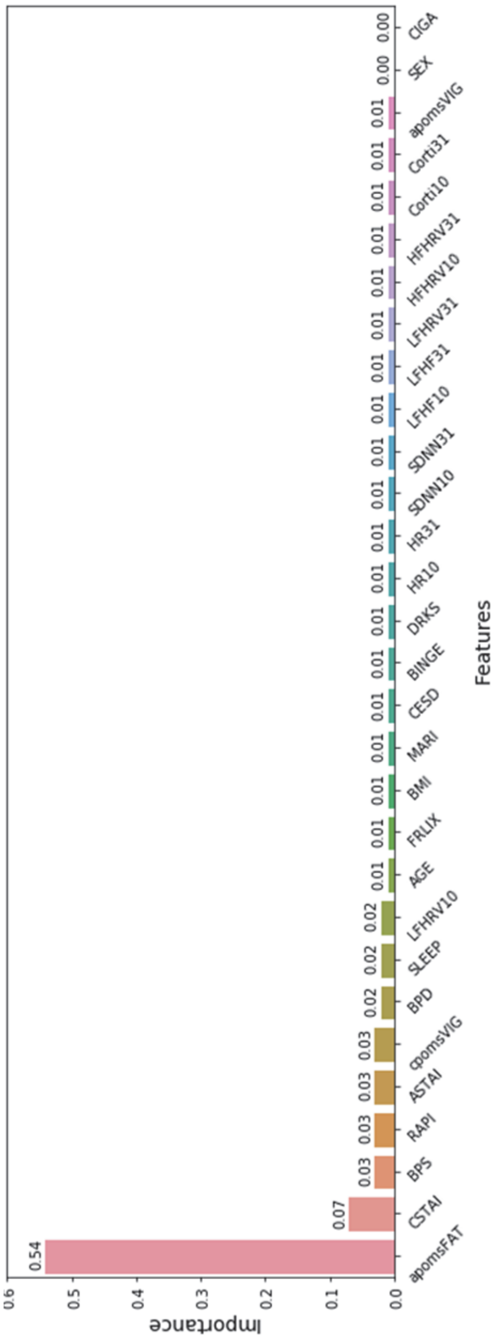

Figure 4:
Measures of variable (feature) importance in predicting post-experiment fatigue

Table 2:
Predictors with the Highest Correlation vs. Importance in Descending Order

	Correlation			Importance	
1	apomsFAT^{S}	0.73	1	apomsFAT^{S}	0.54
2	CSTAI^{S}	0.45	2	CSTAI^{S}	0.07
3	ASTAI^{S}	0.40	3	BPS^{S}	0.03
4	CESD^{BC}	0.33	4	RAPI^{RF}	0.03
5	cpomsVIG^{S}	-0.30	5	ASTAI^{S}	0.03
6	BPS^{S}	-0.21	6	cpomsVIG^{S}	0.03
7	BPD^{S}	-0.20	7	BPD^{S}	0.02
8	LFHF10^{BC}	0.16	8	SLEEP^{RF}	0.02
9	HR31^{BC}	0.13	9	LFHRV10^{RF}	0.02
10	HFHRV10	-0.13	10	AGE	0.01
11	HR10	0.12	11	FRLIX	0.01
12	FRLIX	0.12	12	BMI	0.01
13	MARI	0.11	13	MARI	0.01
14	apomsVIG	-0.11	14	CESD	0.01
15	SDNN10	-0.11	15	BINGE	0.01
16	SEX	-0.10	16	DRKS	0.01
17	AGE	-0.10	17	HR10	0.01
18	SLEEP	-0.08	18	HR31	0.01
19	BMI	-0.08	19	SDNN10	0.01
20	HFHRV31	-0.07	20	SDNN31	0.01
21	LFHF31	0.06	21	LFHF10	0.01
22	RAPI	0.05	22	LFHF31	0.01
23	DRKS	0.04	23	LFHRV31	0.01
24	BINGE	0.04	24	HFHRV10	0.01
25	CIGA	-0.04	25	HFHRV31	0.01
26	Corti10	0.04	26	Corti10	0.01
27	Corti31	0.03	27	Corti31	0.01
28	LFHRV31	0.02	28	apomsVIG	0.01
29	LFHRV10	0.01	29	SEX	0.00
30	SDNN31	0.00	30	CIGA	0.00

Notes. ^{S} = These variables were correlated with the outcome variable and important in a random forest regression analysis. ^{RF} = These variables were not strongly correlated with the outcome variable, but had greater importance in reducing prediction error in a random forest regression analysis. ^{BC} = These variables had a strong bivariate correlation with the outcome variable, but were not important in a random forest regression analysis.

Figure 4 shows that the level of fatigue that participants reported at the outset of the experiment was the most dominant variable (54%) in predicting post-experiment fatigue, which was not surprising. Feeling anxious post experiment (7%) was also quite important in correctly predicting post-experiment fatigue. Beyond these two variables, systolic blood pressure, alcohol-related problems, pre-experiment anxiety, less vigor post experiment, diastolic blood pressure, sleep, and LF HRV were important in adding accuracy in prediction. Other lifestyle variables (BMI and physical activity), alcohol and drug use behaviors (alcohol quantity and cigarette and marijuana use frequency), demographic variables (age and sex), and any other physiological or biological markers were less important, with each contributing 1% or less toward prediction.

Interestingly, variables that would have been missed based on a bivariate analysis (see Table 2) emerged important in the analysis. They were alcohol-related problems, sleep, and LF HRV at baseline. This result suggests that higher-order interactions involving alcohol-related problems, sleep, and LF HRV may exist to predict post-experiment fatigue. What may be noteworthy is that although alcohol-related problems was highly correlated with other alcohol and drug use behaviors and weakly correlated with post-experiment fatigue (see the correlation matrix in a heat map in Figure 1 and also Table 2), it is the alcohol-related problems, not consumption per se that was predictive of post-experiment fatigue. Similarly, LF HRV is thought to capture one's adaptive reactions to a stressful "fight-or-flight" situation in which one needs to mobilize energy to effectively respond. Although LF HRV was highly correlated with other HRV indices, no other indices emerged as important variables. Along with sleep, which is increasingly seen implicated in emotion regulation (Palmer & Alfano, 2017), alcohol-related problems and LF HRV at baseline may be critical. State anxiety at both time points, despite its clear dependency across time within persons, were also important in improving prediction accuracy. Anticipatory anxiety pre experiment, as well as not being able to return to normalcy quickly post experiment may contribute to feeling fatigued. Systolic and diastolic blood pressure readings, which were independently measured in the lab, were important as well. Age and sex and other variables had little to no contributions.

Predicting pre-experiment fatigue

We subsequently analyzed which of the baseline measures would predict pre-experiment fatigue. We chose 200 tree estimators and used the same options (auto, 10-fold cross validation). The model explained 89% of the variance, with the OOB R^2 of 0.20, the mean RMSE = 0.72 (*SD* = .18). Figure 5 shows that pre-experiment anxiety, sleep, HF HRV at baseline, physical activity, depressive symptoms, mean HR at baseline, cortisol at baseline, systolic blood pressure, BMI, alcohol-related problems, and SDNN at baseline were important.

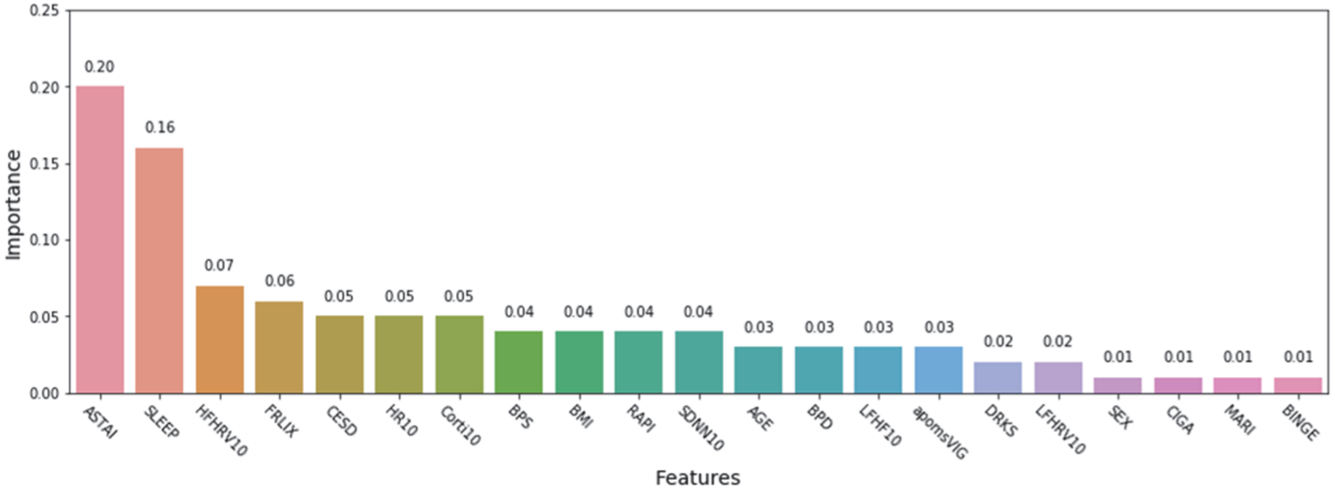

Figure 5:
Measures of baseline variable (feature) importance in predicting pre-experiment fatigue

Ordinary least square regression and classification and regression tree

To provide some context for the results of the random forest analysis, we also conducted an ordinary least square (OLS) regression analysis with all predictors included in a model. The R^2 and adjusted R^2 estimates were .70 and .64, respectively. Out of the 30 predictors, the following four were statistically significant at alpha level 0.05: pre-experiment fatigue (standardized coefficient = 0.68, t = 12.74), pre- and post-experiment vigor (0.33, t = 5.11; -0.37, t = -5.47, respectively), and cigarette smoking frequency in the past month (-0.16, t = -2.76). The variance influence factor (VIF) ranged from 1.21 to 18.12. In sum, the random forest regression analysis had a substantial gain in prediction accuracy, compared to OLS regression, which also showed a different set of variables with importance.

We also ran a CART model. The CART analysis pointed to the following nine variables as relatively important: Pre-experiment fatigue, pre- and post-experiment anxiety, post-experiment vigor, systolic blood pressure at baseline, physical activity, depressive symptoms, alcohol-related problems, diastolic blood pressure at baseline. The CART is known to overfit and to be more susceptible to model instability and greater prediction error, compared to the random forest regression model (Vasconcelos, 2017).

Discussion

The current study utilized a random forest regression analysis to identify variables that help explain individual differences in adaptability to a mildly uncomfortable situation (i.e., post-experiment fatigue). We also examined pre-experiment fatigue in a separate analysis using only trait-like variables and baseline physiological measures. Substantively, fatigue among healthy young adults has not been frequently studied. However, fatigue may sensitively reflect one's ongoing adaptation, or lack of it, to changing environmental demands and stress. Although many physiological indices used in this study are not known to sensitively differentiate healthy young adults across the subtle health risk spectrum, the current study suggests that discoverable patterns of disruptions may exist across multiple systems domains, including the ANS functioning, mood regulation, and self-regulation.

In an earlier study of college drinkers, greater SDNN and lower HF HRV at baseline were conceptualized as general indicators of health, which were predicted by greater exercise and greater alcohol consumption, respectively, from OLS regression models that included BMI, alcohol consumption, cigarette use, exercise, sleep, and other covariates, while none predicted LF HRV (Udo et al., 2013). The current study suggests that a simple assessment approach involving mean HR, blood pressure, current mood, and general health behaviors may suffice to detect pre- and post-experiment fatigue, which may be conceptualized as indicators of an individual's momentary health condition and adaptability and robustness, respectively. Sleep and alcohol-related problems, in particular, had low bivariate correlations with the outcome variable but were important. Blood pressure readings were also important not only for predicting post-experiment fatigue but also pre-experiment fatigue. These findings may suggest that variables like alcohol-related problems and blood pressure may be used to sensitively differentiate individuals at multiple splits and in connection with other variables. With emerging wearable devices and smartphones for monitoring blood pressure, heart rate, exercise, sleep, and alcohol use, the finding from this study is promising for developing a useful screening algorithm for general health and adaptability. In addition, it is increasingly seen as a major gap to better understand and ultimately lower cravings prior to, or during, high risk situations by using an ecological momentary assessment approach (Brannon, Cushing, Crick, & Mitchell, 2016; Singh & Björling, 2019). Computationally intensive approaches may be helpful for advancing the health science field that has been rapidly accumulating high dimensional data (e.g., fMRI) or big data (e.g., electronic health record data).

It is also important to describe how the random forest regression used in this study is different from more common OLS regression or CART. In random forests, methodologically, individual importance estimates of correlated or dependent predictors tend to be lower because a split on one variable will reduce the likelihood of the other variable being subsequently selected when growing trees in random forests (Breiman, 2001; van der Meer et al., 2017). The fact that many of the indices or features (i.e., state anxiety at both time points, systolic and diastolic blood pressure, alcohol-related problems, LF HRV at baseline, POMS vigor at conclusion) were either observed at two time points or were highly correlated with other variables, yet emerged as important may suggest that these variables have unique and complex relationships with the rest of the variables that are hard to detect in a standard OLS regression analysis that focuses on additive linear effects. In contrast to a single tree decision model, the random forest model utilized computationally intensive bootstrapping and bagging so that the importance of each variable can be accurately assessed.

We used Scikit-Learn, a free software machine learning library for the Python programming language for this analysis. However, other options also exist. There is a 'randomForest' package (Liaw, & Wiener, 2002) for analysis in R (R Core Team, 2019). There is WEKA (Frank, Hall, & Witten, 2016), a stand-alone software written in Java for a suite of machine learning algorithms. SAS Institute (2016) has a random forest regression and classification routine called 'proc hpforest.' We tested Scikit-Learn in Python (version 3.7, Python Core Team, 2019), proc hpforest in SAS (version 9.4), and Weka (version 3.8.3) before choosing to go with Scikit-Learn for Python. A recent comparative analysis of R, SAS, and Python for random forests showed that the results were very similar in terms of error rates. However, the routines by Python and SAS captured variable importance better than R. In addition, SAS proc hpforest was considerably slower than R or Python (Soifua, 2018). Based on our experience analyzing the data reported, random forest regression and classification can be relatively easily implemented for research applications. However, more advances in methods would be welcomed. For example, there is no easy way to calculate confidence intervals surrounding the key metric for features (i.e., relative importance) or to better probe the complex relationships. Currently, there is no good way to graphically display random forest results. Hence, random forests are sometimes called a "black box" model (Strobl et al., 2008). With methodological advances, a follow-up investigation on the mechanistic associations may be needed to fully leverage the findings from this sensitive approach. Nonetheless, this is a promising alternative to the existing approaches in major ways and should be explored more in many scientific fields, including psychology.

Findings from this study should be interpreted with cautions. First, this study reports data from a relatively higher percentage of Asian young adults, many of them were college students and graduate students at a state university in the Northeastern United States. Therefore, the study's generalizability may be limited. Second, due to the study exclusion criteria, we removed those whose blood pressure readings exceeded 140/90, which restricted the range of blood pressure data. The restricted measurement range typically weakens the true magnitude of the relationships. Similarly, we also excluded those who were more than 20% over- or under-weight from the ideal for gender, height, and body frame based on the Metropolitan Life Height–Weight Table (Metropolitan Life Insurance Company, 1983). Therefore, any effects of BMI or physical activity on fatigue may have been attenuated. Third, this study sample was relatively small and heterogeneous, which may explain the lower prediction for OOB samples, compared to prediction for training data sets. Fourth, random forest regression results are difficult to interpret. Therefore, we could not delineate underlying complex relationships among features. An additional work with the help of new methodological tools may be needed to shed light on complex high-order interactions that are likely to be present in the data. Finally, although it is reasonable to assume that participating in an experimental study was mildly stressful, we did not directly assess how stressful participants felt about their study participation experience.

Having discussed the caveats of this study, we now highlight that this was one of the earliest research applications utilizing a machine learning algorithm. We predicted post-experiment fatigue by using a wide range of psychosocial, lifestyle, and physiological indicators. We conclude that a number of health-promotive lifestyle factors (sleep, physical activity, and BMI), physiological indicators (blood pressure, mean HR, HF HRV, cortisol, SDNN), and psychological functioning variables (state anxiety, depression, and alcohol-related problems) were involved in predicting pre-experiment fatigue as a general health state. However, post-experiment fatigue as a response to changes in the environment, after setting aside its earlier state, was predicted mostly by state anxiety, blood pressure, alcohol-related problems, sleep, post-experiment vigor and LF HRV at baseline. The results may suggest that feeling easily fatigued following a mildly stressful situation may signal disruptions involving multiple related systems domains, including the ANS functioning, mood regulation, and self-regulation. We also conclude that the promise of new computing advances is intriguing although more work is needed.

Acknowledgement

This study was supported by Colgate Palmolive and in part by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) of the National Institutes of Health under award number R01 AA019511. We would like to thank Patricia Buckendahl, Anne E. Ray, and other members of the Sensation and Emotion Lab for their help with data collection.

References

American Statistical Association (2018). International Prize in Statistics awarded to Bradley Efron. https://www.amstat.org/asa/News/International-Prize-in-Statistics-Awarded-to-Bradley-Efron.aspx

Appelhans, B. M., & Luecken, L. J. (2006). Heart rate variability as an index of regulated emotional responding. *Review of General Psychology, 10*(3), 229-240. https://doi.org/10.1037/1089-2680.10.3.229

Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis, 52*(4), 2249-2260. doi:https://doi.org/10.1016/j.csda.2007.08.015

Berntson, G. G., Bigger, J. T., Eckberg, D. L., Grossman, P., Kaufmann, P. H., Malik, M., et al. (1997). Heart rate variability: Origins, methods and interpretive caveats. *Psychophysiology, 34*, 623-648. https://doi.org/10.1111/j.1469-8986.1997.tb02140.x

Brannon, E. E., Cushing, C. C., Crick, C. J., & Mitchell, T. B. (2016). The promise of wearable sensors and ecological momentary assessment measures for dynamical systems modeling in adolescents: a feasibility and acceptability study. *Translational Behavioral Medicine, 6*(4), 558-565. https://doi.org/10.1007/s13142-016-0442-4

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cochran, J. J. (2018). What is the bootstrap? *Significance, 15, 6*. https://www.significancemagazine.com/science/608-what-is-the-bootstrap

Collins, R. L., Parks, G. A., & Marlatt, G. A. (1985). Social determinants of alcohol consumption: The effects of social interaction and model status on the self-administration of alcohol. *Journal of Consulting and Clinical Psychology, 53*, 189–200.

Dankowski, T., & Ziegler, A. (2016). Calibrating random forests for probability estimation. *Statistics in Medicine, 35*(22), 3949-3960. https://doi.org/10.1002/sim.6959

Dobler, V. B. et al. (2019). Disaggregating physiological components of cortisol output: A novel approach to cortisol analysis in a clinical sample – A proof-of-principle study. *Neurobiology of Stress, 10*, 100153. https://doi.org/10.1016/j.ynstr.2019.100153

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. *The Annals of Statistis, 7*(1), 1-26. https://doi.org/10.1214/aos/1176344552

Frank, E. et al. (2010). Weka - A machine learning workbench for data mining. In O. Maimon & L. Rokach (Eds.), *Data mining and knowledge discovery handbook* (pp. 1269-1277). Boston, MA: Springer.

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* Boston: O’Reilly.

Grömping, U. (2009). Variable importance assessment in regression: Linear regression versus random forest. *The American Statistician, 63*(4), 308-319. doi:10.1198/tast.2009.08199

Gunnar, M. R., & Davis, E. P. (2003). Stress and emotion in early childhood. In R. M. Lerner, M. A. Easterbrooks, & J. Mistry (Eds.), *Handbook of psychology: Developmental psychology* (pp. 113-134). New York: Wiley.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods, 21*(4), 447-457. https://doi.org/10.1037/met0000120

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician, 69*(4), 371-386. https://doi.org/ 10.1080/00031305.2015.1089789

Jennings, J. R., Kamarck, T., Stewart, C., Eddy, M., & Johnson, P. (1992). Alternate cardiovascular baseline assessment techniques: vanilla or resting baseline. *Psychophysiology*, *29*, 742-750. https://doi.org/10.1111/j.1469-8986.1992.tb02052.x

Kokubun, K., Nemoto, K., Oka, H., Fukuda, H., Yamakawa, Y., & Watanabe, Y. (2018). Association of fatigue and stress with gray matter volume. *Frontiers in Behavioral Neuroscience, 12*(154). https://doi.org/10.3389/fnbeh.2018.00154

Koopman, J., Howe, M. Hollenbeck, J. R., & Sin, H.-P. (2015). Small sample mediation testing: Misplaced confidence in bootstrapped confidence intervals. *Journal of Applied Psychology, 100(1),* 194-202. https://doi.org/10.1037/a0036635

Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R News, 2*(3), 18-22.

McNair, D. M., Lorr, M., & Droppleman, L. F. (1992). *Profile of Mood States Manual*. North Tonawanda, NY: Multi-Health Systems.

Metropolitan Life Insurance Company. (1983). The metropolitan life height-weight table. *Statistical Bulletin, 64,* 2–9.

Mucci, N., Giorgi, G., De Pasquale Ceratti, S., Fiz-Pérez, J., Mucci, F., & Arcangeli, G. (2016). Anxiety, stress-related factors, and blood pressure in young adults. *Frontiers in Psychology, 7*(1682). https://doi.org/10.3389/fpsyg.2016.01682

Mun, E.-Y., von Eye, A., Bates, M. E., & Vaschillo, E. G. (2008). Finding groups using model-based cluster analysis: Heterogeneous emotional self-regulatory processes and heavy alcohol use risk. *Developmental Psychology, 44*(2), 481-495. https://doi.org/10.1037/0012-1649.44.2.481

Mun, E.-Y., Windle, M., & Schainker, L. M. (2008). A model-based cluster analysis approach to adolescent problem behaviors and young adult outcomes. *Development and Psychopathology, 20*(1), 291-318. https://doi.org/10.1017/S095457940800014X

National Institutes of Health (2018). NIH strategic plan for data science. https://datascience.nih.gov/sites/default/files/NIH_Strategic_Plan_for_Data_Science_Final_508.pdf.

Palmer, C. A., & Alfano, C. A. (2017). Sleep and emotion regulation: An organizing, integrative review. *Sleep Medicine Reviews, 31*, 6-16. https://doi.org/10.1016/j.smrv.2015.12.006

Python Core Team (2019). Python: A dynamic, open source programming language. Python Software Foundation. https://www.python.org/.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

SAS Institute Inc. (2016). *SAS Enterprise miner 14.2: High-performance procedures.* Cary, NC: SAS Institute Inc.

Scheidel, C. (2018). *Be aware of bias in random forest variable importance metrics*. https://blog.methodsconsultants.com/posts/be-aware-of-bias-in-rf-variable-importance-metrics/

Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health, 5*, 258. https://doi.org/10.3389/fpubh.2017.00258

Singh, N. B., & Björling, E. A. (2019). A review of EMA assessment period reporting for mood variables in substance use research: Expanding existing EMA guidelines. *Addictive Behaviors, 94*, 133-146. https://doi.org/10.1016/j.addbeh.2019.01.033

Sloan, R. P. et al. (1994). Effect of mental stress throughout the day on cardiac autonomic control. *Biological Psychology, 37*(2), 89-99. https://doi.org/10.1016/0301-0511(94) 90024-8

Soifua, B. (2018). *A comparison of R, SAS, and Python implementations of random forests*. All Graduate Plan B and other Reports. 1268. https://digitalcommons.usu.edu/grad reports/1268

Spielberger, C. D., Gorsuch, R. C., Lushene, R. E., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Steingrimsson, J. A., Diao, L., & Strawderman, R. L. (2019). Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association, 114*(525), 370-383. https://doi.org/10.1080/01621459.2017.1407775

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*(1), 307. https://doi.org/10.1186/1471-2105-9-307

Tiesinga, L. J., Dassen, T. W. N., Halfens, R. J. G., & van den Heuvel, W. J. A. (1999). Factors related to fatigue; priority of interventions to reduce or eliminate fatigue and the exploration of a multidisciplinary research model for further study of fatigue. *International Journal of Nursing Studies, 36*(4), 265-280. https://doi.org/10.1016/S0020-7489(99)000 22-X

Travis, F. et al. (2018). Effect of meditation on psychological distress and brain functioning: A randomized controlled study. *Brain and Cognition, 125*, 100-105. https://doi.org/10.1016/j.bandc.2018.03.011

Udo, T., Mun, E.-Y., Buckman, J. F., Vaschillo, E. G., Vaschillo, B., & Bates, M. E. (2013). Potential side effects of unhealthy lifestyle choices and health risks on basal and reactive heart rate variability in college drinkers. *Journal of Studies on Alcohol and Drugs, 74*(5), 787-796. https://doi.org/10.15288/jsad.2013.74.787

United States Department of Health and Human Services (USDHHS, 2009). Summary health Statistics for U.S. children: National Health Interview Survey, 2007: Data from the National Health Interview Survey. *Vital and Health Statistics, 10* (239), 1-88.

White, H. R., & Labouvie, E. W. (1989). Towards the assessment of adolescent problem drinking. *Journal of Studies on Alcohol, 50,* 30–37.

White, H. R., & Labouvie, E. W. (2000). Longitudinal trends in problem drinking as measured by the Rutgers Alcohol Problem Index. *Alcoholism: Clinical and Experimental Research, 24,* 76A.

Vasconcelos, G. (2017). How random forests improve simple regression trees? Accessed at https://insightr.wordpress.com/2017/09/23/how-random-forests-improve-simple-regression-trees/

van der Meer, D. et al. (2017). Predicting attention-deficit/hyperactivity disorder severity from psychosocial stress and stress-response genes: a random forest regression approach. *Translational Psychiatry, 7*, e1145. https://doi.org/10.1038/tp.2017.114