Published in final edited form as:
Biol Psychol. 2022 April ; 170: 108290. doi:10.1016/j.biopsycho.2022.108290.

Objective underpinnings of self-reported sleep quality in middle-aged and older adults: the importance of N2 and wakefulness

Renske Lok[1], **Dwijen Chawra**[1], **Flora Hon**[1,2], **Michelle Ha**[3], **Kate A. Kaplan**[1], **Jamie M. Zeitzer**[1,4,*]

[1]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford CA 94305

[2]College of Literature, Science, and the Arts, University of Michigan, Ann Arbor MI 48109

[3]Department of Mathematics and Statistics, San Jose State University, San Jose CA 95112

[4]Mental Illness Research Education and Clinical Center, VA Palo Alto Health Care System, Palo Alto CA 94304

Abstract

Study Objectives: The measurable aspects of brain function (polysomnography, PSG) that are correlated with sleep satisfaction are poorly understood. Using recent developments in automated sleep scoring, which remove the within- and between-rater error associated with human scoring, we examine whether PSG measures are associated with sleep satisfaction.

Design and Setting: A single night of PSG data was compared to contemporaneously collected measures of sleep satisfaction with Random Forest regressions. Whole and partial night PSG data were scored using a novel machine learning algorithm.

Participants: Community-dwelling adults (N=3,165) who participated in the Sleep Heart Health Study.

Interventions: None

Measurements and Results: Models explained 30% of sleep depth and 27% of sleep restfulness, with a similar top four predictors: minutes of N2 sleep, sleep efficiency, age, and minutes of wake after sleep onset (WASO). With increasing self-reported sleep quality, there was a progressive increase in N2 and decrease in WASO of similar magnitude, without systematic changes in N1, N3 or REM sleep. In comparing those with the best and worst self-reported sleep satisfaction, there was a range of approximately 30 minutes more N2, 30 minutes less WASO, an improvement of sleep efficiency of 7-8%, and an age span of 3-5 years. Examination of sleep most

[*]Correspondence: jzeitzer@stanford.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Disclosure statement.
Financial Disclosure: none.
Non-financial Disclosure: none

proximal to morning awakening revealed no greater explanatory power than the whole-night data set.

Conclusions: Higher N2 and concomitant lower wake is associated with improved sleep satisfaction. Interventions that specifically target these may be suitable for improving the self-reported sleep experience.

Keywords

Machine learning; sleep; sleep quality; polysomnography; human; adult

Introduction.

What is good sleep? To paraphrase U.S. Supreme Court Justice Potter Stewart, “I know it when I see it.” While this answer may be unsatisfying, it engenders an important question: what are the aspects of physiology that underlie our self-reported experience of sleep? Beyond an esoteric concept, the self-reported sleep experience is relevant for multiple health outcomes such as mortality (1), metabolic syndrome (2), diabetes (3), hypertension (4), coronary heart disease (5), schizophrenia (6), autism (7) and depression (8). An understanding of the physiologic underpinnings of self-reported sleep quality not only allow us to understand the mechanistic relationship between sleep and psychiatric disease, it could also facilitate treatments targeted to the specific aspects of sleep that are linked to an improved sleep experience.

One aspect of brain physiology that is frequently captured during studies of sleep is that of polysomnography, which is used to parse 30-second segments of brain wave (electroencephalography, EEG), muscle (electromyography, EMG) and eye movement (electro-oculography, EOG) data into ‘stages’ of sleep. These stages are divided into rapid eye movement (REM) and non-REM sleep (NREM), which is further divided into three separate categories (N1, N2, N3). Previous studies comparing PSG with self-reported sleep quality have identified each of the stages of NREM sleep (N1, N2, N3), wake after sleep onset (WASO), transitions between sleep and wake, and overall sleep efficiency (SE) (9–16) as being important contributors to the subjective experience of sleep. Many of these studies, however, relied on small numbers of individuals who were either good sleepers or who were clinically diagnosed with insomnia. We have previously published on the association between self-reported sleep quality and polysomnography (PSG) measures of sleep in multiple, large, community-based cohorts of middle to older aged adults (17–19). While these studies indicated that SE, WASO, and total sleep time (TST) were important physiologic correlates of self-reported sleep quality, the models explained relatively little variance (≤15%).

Since these initial publications, advances in both sleep analyses and machine learning warrant revisiting the relationship between self-reported sleep quality and its physiologic correlates. Classification into sleep stages is typically done by experts trained to detect specific patterns in the EOG, EMG, and EEG signals and to match these patterns to manualized standards (20,21). In scoring the PSG, there are well-described, substantive interindividual differences among sleep experts, as well as intraindividual inconsistencies, that can complicate analyses of cross-sectional studies (22,23). As such, there have been many efforts in recent years to use machine learning-based automated scoring techniques to, at the very least, eliminate both within- and between-scorer bias and, ideally, improve scoring accuracy. In one such iteration of this approach, labels (wake, N1, N2, N3, REM) are provided for each 15-s epoch of sleep, as well as the probability of each state within each 15-s epoch (24,25). In other words, for each 15-s epoch, this approach assesses how closely EEG, EOG, and EMG signals match the idealized pattern for the state as determined by the model. Information on probability of state matching could be helpful in understanding the confidence of sleep staging, which could be secondary to temporal (multiple states occurring within the 15-s epoch) or spatial (multiple states occurring in cortical regions contiguous with the recording electrode) integration in the EEG signal (26).

Another possible reason why PSG variables in prior studies have not predicted self-reported sleep quality well is that these studies often relied on whole-night measures of sleep. It is possible that the self-reported experience of sleep quality is dependent on the sleep occurring more proximal to the final awakening. As such, the latter part of the PSG might be a better predictor of self-reported sleep quality. The goals of this study are to revisit the relationship between PSG variables and self-reported sleep quality and to determine whether this relationship changes based on the part of the night from which the data were obtained. To address these goals, data from the Sleep Heart Health Study were examined with machine learning regression models meant to reduce the chance of overfitting the predictive data.

Methods and Materials.

Original data were collected in the Sleep Heart Health Study, a multi-center clinical cohort originally designed to examine the cardiovascular consequences of disrupted breathing during sleep. Data were collected between November 1, 1995 and January 31, 1998. Of the 6,441 participants in the Sleep Heart Health Study, 5,804 had overnight sleep data available for analysis. Data were excluded if either of the self-reported sleep quality assessments (explained below) were missing (n=376) or if there were issues with the quality (e.g., poor signal) or quantity (e.g., lack of a complete night) of the PSG (n=2,263), leaving a final sample of 3,165 individuals. Complete information about the original study methods and design is available elsewhere (27); information specific to these analyses is presented below. Data were accessed from the National Sleep Research Resource (www.sleepdata.org, v. 0.15.0) (27,28). Participant consent was obtained by the individual institutions involved in the Sleep Heart Health Study.

Self-reported sleep quality assessment.

The primary outcome variables of interest in these analyses are ‘rest10’ and ‘Itdp10’. Both variables concern the self-reported quality of sleep and were asked in the morning survey immediately following an overnight PSG. The former (rest10) asks participants to self-rate the quality of their sleep based on the restfulness of their sleep; two anchors [restless (1), restful (5)] were used on a five-point Likert-like scale. The latter (Itdp10) asks participants to self-rate the quality of their sleep based on the depth of their sleep; two anchors [light (1), deep (5)] were used on a five-point Likert-like scale.

Polysomnography.

PSG was recorded in the participant’s home using a PS-2 system (Compumedics, Abbotsford, Australia) (27). These data were downloaded as original, unscored PSG data (.edf files). We rescored these PSG data with a machine learning-based neural network algorithm (https://github.com/Stanford-STAGES/stanford-stages) on a high-performance computing cluster at Stanford University (Sherlock) (24). This algorithm was initially trained on PSG records from thousands of individuals with a variety of sleep pathologies (25). The algorithm scores data in 15-s epochs, rather than the typical 30-s epochs, and provides labels of N1, N2, N3, REM, and Wake for each 15-s epoch. In addition to the number of minutes spent in each stage, we also calculated the number of transitions between any sleep stage and wake, the number of transitions between N3 and either N1 or N2, sleep latency (time from lights out to sleep onset, defined as the first occurrence of N2 or three consecutive stages of N1), and sleep efficiency (total sleep time divided by time in bed). In addition to these traditional metrics, the algorithm also describes the relative probability of each stage (i.e., closeness of matching the pattern for that stage) for each epoch. For example, a given epoch scored as N3 might have N1=0.05, N2=0.03, N3=0.85, REM=0.04, Wake=0.03, indicating that the electrophysiologic pattern within that epoch very closely resembles that of N3 sleep, with very low matching patterns of N1, N2, REM, or Wake. From these data, we derived two additional novel measures, the average probability of each stage (e.g., for each epoch defined as N1, what is the average probability of N1 in these epochs) and the adjusted wake amount (i.e., average probability score of wake in epochs scored as wake, multiplied by the total amount of wake after sleep onset). For the purposes of this manuscript, these data are referred to as “auto-scored”.

To examine whether PSG data obtained closer to wake time held more relevance to self-reported sleep scores, whole night auto-scored data were parsed into fragments based on time relative to wake. Data from 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 220, and 240 minutes from wake time (e.g., 80 minutes would be the final 80 minutes of PSG data counting backwards from wake time) were derived and analyzed as described below. This sequential analysis was limited to the last 240 minutes (4 hours) of sleep as durations longer than this would have begun to exclude individuals who had shorter sleep durations.

To compare the results of the auto-scoring, we also examined the previously hand-scored PSG data. In the original Sleep Heart Health Study, expert polysomnographic technicians scored 30-s epochs of PSG data as stages of S1 (corresponding to a current designation of N1), S2 (corresponding to a current designation of N2), S3 and S4 (corresponding to a current designation of N3), REM, and wake. Determination of sleep stage was done according to standard Rechtschaffen and Kales criteria (29). Nightly amounts of each sleep stage (N1, N2, N3, REM, wake) were calculated. For the purposes of this manuscript, these data are referred to as “hand-scored”. From the PSG data, we also used the overall respiratory disturbance index (RDI), calculated as the count of all apneas plus hypopneas with at least a 4% oxygen desaturation, divided by hours of total sleep time. Auto-scored and hand-scored data were compared with paired t-tests and effect sizes were calculated by the Standardized Mean Difference (SMD). SMD values correspond to small (0.2), medium (0.5), and large (0.8) effect sizes (30).

Other Predictor Variables.

In addition to the sleep data, a variety of other data potentially related to sleep quality were also included in the models. These include demographic variables [gender (male or female), race (Hispanic/Latino or non-Hispanic/Latino), age, ethnicity, education level (grouped as <10 years, 11–15 years, 16-20 years, >20 years), marital status (married, divorced/separated, widowed, never married, or unknown)], anthropometric variables [body mass index (kg/m^2), waist circumference (cm)], trait-like sleep measures [Epworth Sleepiness Scale (31), habitual sleep duration], variables that captured behavior during the four hours prior to the overnight sleep study [number of caffeinated drinks, number of alcoholic drinks, number of nicotine products], typical medication use [antidepressant or benzodiazepine use within two weeks of study and use of sleeping pills at least one day per week], self-rated health, overall cognitive health [Mental Component Scale standardized score from the SF-36, MCS (32)] physical health [Physical Component Scale standardized score from the SF-36, PCS (32)] and self-rated emotional status [feeling calm and peaceful during previous four weeks (rated as all, most, some, a little, or none of the time) and stressfulness of previous day (typical, less, or more)]. In this sample, 1.1% of the non-sleep data were missing, with the greatest amount (6.9%) being MCS and PCS scores. Missing data were imputed with AmeliaView 1.7.3(33), using the fifth iteration and bounds where appropriate.

Machine learning.

To predict self-reported sleep quality, a Random Forest (RF) regression analysis (classification algorithm function, RandomForestRegressor in sklearn v. 0.24.1; run in Python v. 3.7.10) was used. RF is a machine learning algorithm that can be used for regression analysis particularly when overfitting is a concern (34). It deals well with non-linearity, is not heavily impacted by noise, and is robust to inclusion of both categorical and continuous variables (35). Data were randomly split into sets of 75% for training (developing the model) and 25% for testing the model. The Python function RandomizedSearchCV class (sklearn) was used to tune model hyperparameters: the number of estimators, maximum depth of a tree, and minimum samples required at leaf node. The function GridSearchCV was used to automatically determine the maximum features per split. Similar hyperparameters were determined for each of the models (complete data set and iterative fragments of the night). We therefore used a weighted average of all parameters, using the r^2 of the test set model for weighting, for each of the models. Variables that contributed at least 5% of the explained variance were further examined for trend per self-reported sleep quality score, using a Cuzick trend test set for Wilcoxon rank (package “PMCMRplus” in R, version 1.4.1103). To test for monotonic increases or decreases, the ‘is.unsorted’ function in R was used.

Results.

The sample (n=3,165) was about half female, mostly White, middle- and older aged adults (Table 1). Sleep variables were in the ranges to be expected in a community-based cohort (Table 1). Following the overnight PSG, most individuals rated their sleep as moderately deep and moderately restful (Table 1). These two aspects of self-reported sleep quality were correlated (Spearman rho=0.68, p<0.001) (Figure 1). When people reported having the most restful sleep, they usually also reported having the deepest sleep, and vice versa (highest correspondence in the 1 and 5 categories). People reporting moderate depth or restfulness (categories 2-4) had greater divergence.

RF regressions were used to examine the relationship between PSG predictor variables (Table 1) and both self-reported sleep depth and restfulness. Hyperparameters tuning resulted in 2000 estimated trees, a maximum tree depth of 10, and a minimum of 2 samples for each leaf node. The RF models for both outcomes explained similar amounts of variance (29.5% for self-reported depth, 26.8% for self-reported restfulness) and had a similar top four predictors: minutes of N2, sleep efficiency, age, and minutes of WASO (Figure 2). These top four predictors captured 28% and 26% of the relative model variance for self-reported restfulness and self-reported depth, respectively.

While the RF are useful for understanding the combined prediction of the input variables, we also used these models as feature selectors. To contextualize the contribution of the different PSG features selected by the RF models, the four variables (N2, WASO, SE, age) that contributed at least 5% of the explained variance were further examined in isolation (Figure 3). With each increase (improvement) in one unit of sleep depth, there was an increase in age of 0.79 years (Figure 3A; Cuzick trend test, z=4.22, p<0.0001) and with each increase (improvement) in one unit of sleep restfulness, there was an increase in age of 1.2 years (Figure 3B; z=7.12, p<0.0001). This leads to a relatively narrow span of 3.2 and 4.8 years for the range of subjective depth and restfulness scores, respectively. However, the change in age associated with sleep depth and restfulness did not follow a monotonic pattern. Each increase in one unit of sleep depth was associated with a monotonic increase in N2 by 7.6 minutes (Figure 3C; z=8.90, p<0.0001), a monotonic 2.0% increase in sleep efficiency (Figure 3E; z=9.63, p<0.0001), and a monotonic decrease in WASO by 7.2 minutes (Figure 3G; z=−8.86, p<0.0001). With each increase in one unit of restfulness (improvement), there was a monotonic increase in N2 by 6.7 minutes (Figure 3D; z=8.43, p<0.0001), a monotonic 1.7% increase in sleep efficiency (Figure 3F; z=8.36, p<0.0001), and a monotonic decrease in WASO by 6.2 minutes (Figure 3H; z=−8.04, p<0.0001). These translate to a range of sleep depth that spans a 30-minute difference in N2, 33-minute difference in WASO, and 8.0% difference in sleep efficiency and a range of sleep restfulness that spans a 27-minute difference in N2, 25-minute difference in WASO, and 6.8% difference in sleep efficiency.

These changes in N2 and WASO can also be visualized when examining the relative proportion of the night spent in different stages of sleep. With increasing self-reported sleep quality, there is a progressive increase in N2 and decrease in WASO of similar magnitude, without a corresponding systematic change in N1, N3 or REM for both self-reported sleep depth (Figure 4A) and restfulness (Figure 4B).

In considering the observed decrease in WASO with increase self-reported sleep quality, which is accompanied by an increase in N2 and corresponding improvement in SE, there are three ways in which the amount of wake could decrease: (1) a decrease in the number of wake episodes per night, (2) a decrease in the length of the individual wake episodes, (3) or a combination of the two. If there were a decrease in the number of wake episodes per night, we would expect to observe a corresponding progressive decrease in the number of shifts between sleep and wake, which was not an important part of the model. Even though there is a difference in the number of wake episodes by both sleep depth ($F_{(4,3160)}$=5.54, p=1.92e-4; ANOVA) and restfulness ($F_{(4,3160)}$)=2.93, p=0.0196; ANOVA), post hoc analyses indicate that the difference is not progressive. Those who scored a 2 on depth had more wake episodes than those who scored a 3, 4, or 5 (p's<0.005, Tukey; individual test α=0.005), but the remaining scores did not differ from each other. Additionally, there were no significant post hoc differences between scores on the restfulness scale (p's>.02, Tukey; individual test α =0.005). To explore whether there is a difference in the length of wake episodes, the duration of individual wake episodes was plotted as a cumulative probability plot (Figure 5). Most of the wake episodes were brief, with 85.7% being 2 minutes or shorter. In the wake episodes that are longer than 2 minutes, however, there is a progressive shift such that worse self-reported sleep quality is associated more wake episodes of longer duration (rightward shift in the cumulative probability curves, for both degree of depth (Figure 5A) and restfulness (Figure 5B). The percent of wake episodes longer than 2 minutes is greater in those with worse self-reported sleep quality (p's<0.0001, Kruskal-Wallis ANOVA) such that those with the worst sleep quality have a median of 3.4% (depth, Figure 5B) or 2.6% (restfulness, Figure 5A) more of their wake as these longer episodes.

To examine whether PSG data obtained closer to wake time held more relevance to self-reported sleep satisfaction, whole night auto-scored data were parsed into fragments based on time relative to wake; individual RF models were fit to each fragment of the night. Examination of models derived from data fragments from 20, 40, 60, 80, 100, 120, 140, 160, 180, 200, 220, and 240 minutes before wake time (e.g., 80 minutes would be the final 80 minutes of PSG data counting backwards from wake time) indicated no significant or systematic changes in the amount of explained variance (from 18% to 22% for depth and 16% to 22% for restfulness). While there were some changes in feature importance, there were no systematic changes and the top predictors remained similar in all models (Figure 6). Of the top predictor variables for the whole night, both age and SE were stable in the amount of variance explained in each of the models. N2 and WASO, however, were less important in the models including only the end of the night data, as compared to models including most or all of the night of data (Figure 6).

As the scoring of WASO and N2 appear to be important markers of self-reported sleep satisfaction, we compared how the auto-scoring and hand-scoring performed in identifying epochs as WASO and N2. Epoch-by-epoch comparisons were not possible as the two methods use different analysis windows (15 s vs. 30 s in auto- and hand-scoring, respectively), but we could compare whole-night summary statistics. The auto-scored data had more wake (86.8 ± 63.1 min) as compared to hand-scored (57.7 ± 40.9 min) (p<0.001, paired t-test; SMD=−0.53). The auto-scored data also had more N2 (227 ± 61.3 min) as compared to hand-scored (202 ± 53.4 min) (p<0.001, paired t-test; SMD=−0.43). We examined the stability of the auto-scoring by calculating the average probability of matching each state template within each epoch scored as either wake or N2. The 15-s epochs categorized as "wake" by the auto-scoring algorithm had an average probability of matching wake of 84%, with N1 (9%), N2 (4%), N3 (0%), and REM (2%) capturing the rest of the variance. The 15-s epochs categorized as "N2" by the auto-scoring algorithm had an average probability of matching N2 of 84%, with wake (4%), N1 (7%), N3 (4%), and REM (1%), capturing the rest of the variance.

Discussion.

In a large data set of community-dwelling adults, the amount of wake, N2, overall sleep efficiency, and to a lesser extent, age, are important predictors of self-reported sleep quality on a given night. Specifically, an increase in N2 with a concomitant decrease in the duration of wake after sleep onset, possibly through a decrease in the length of longer wake episodes, is associated with better self-reported sleep depth and restfulness. These changes in sleep states are also reflected in the improved self-reported sleep quality associated with an increase in sleep efficiency. There is a range of approximately 30 minutes of extra N2, 30 minutes less WASO, and an improvement of sleep efficiency of 7-8% in comparing those with the best and worst self-reported sleep quality.

Our data are consistent with the hypothesis that substituting N2 for wake results in higher self-reported sleep quality. Self-reported sleep quality had previously been associated with NREM sleep (N1, N2, N3), WASO, transitions between sleep and wake, and overall SE (9–16). These earlier studies relied on fewer participants and results were not consistent among the studies, likely due to the relatively low power of a single objective variable to explain the variance in self-reported sleep quality. Our findings, which validate the previous association of N2 and wake, used Random Forest regression, which uses a bootstrap approach, reducing the likelihood of a spurious finding due to the nature of repeated sampling. This type of approach, however, also limits discovery of subgroups within the data set. Repeated sampling of the same individuals over multiples nights, as was done in some of the previous studies (9,11), will be important moving forward to determine if different individuals consistently respond to the same set of PSG characteristics. While a reduction in WASO being associated with better sleep quality has prima facie validity, *a priori* we would have expected N3, rather than N2, to be associated with better sleep quality. N3 is a deeper stage of NREM sleep (i.e., higher arousal thresholds than N1 or N2) and is often used colloquially in the literature as ‘sleep quality’. N3, however, was not a significant contributor in any of our models. In part, this might be explained by the scoring algorithm used (25), though we did not previously observe evidence of the involvement of N3 when using hand-scored data (17,18). It also may be that as the amount of N3 is driven by the duration of prior wake (36,37), a reduction in WASO cannot be offset by extra N3 sleep. In examining the WASO, it was not the number of episodes of WASO that was associated with subjective sleep quality, but rather the length of the WASO episodes. While not directly studied, this could be due to a quantal amount of time being necessary before being ‘awake’ has a negative impact on subjective sleep quality. Such an effect could be due to minimal recruitment of multiple brain regions into a state of wake or within a single region, but such hypotheses are beyond the scope of the current study.

As the sleep near the end of the night could have greater weight in determining self-reported sleep quality, we parsed the night into 20-minute segments, starting from a participant’s final awakening. We did not observe, however, increased explanatory power in the RF models in the data segments from the end of the night. Indeed, modelling data from these segments offered less explanatory power than whole-night data sets. Specific variables that were important in the model that predicted self-reported sleep quality by the whole night of polysomnography, such as duration of N2, were less important in the models constructed on data from the end of the night. This is not unexpected given the reduced amounts of NREM and increased amounts of REM during this time frame of sleep. No specific variables appeared to have increased importance at the end of the night.

In the auto-scoring algorithm, each 15-s epoch received not only a label (Wake, N1, N2, N3, REM) but also the degree to which the epoch matched each of the five possible states (based on the initial training algorithm) (24). Thus, if the 15-s epoch contained simultaneous states, either due to temporally contiguous states within the scoring window (e.g., 5 s of N1 followed by 10 s of N2) or due to spatial bleeding (e.g., cortical region under C3 in N1, but nearby cortical regions in N2), the matching probability would be distributed among multiple states. We used this information in the RF models to determine whether the degree to which epochs were classified as singular states was of additional importance. We did not, however, find this to be the case. Rather, the absolute amount of a state (e.g., minutes of N2) was more informative for predicting self-reported sleep quality than how well, on average, a given epoch was representative of a specific state (e.g., probability of matching N2).

Our previous work with this same data (with the same exclusion criteria) set yielded substantially less robust results, with only 8-9% of the variance being explained in RF models, as compared to the 27-30% we observe in this study (18). The previous study had identified sleep efficiency, age, and WASO as top predictors, as does the current study, but also identified total sleep time as important, which was not a highly rated predictor in the current analysis. In our current analysis, N2 was an important predictor, while it was not in the previous study. The disparity between the studies is likely due to two issues. The first is that in this iteration, we tuned the hyperparameters of the models prior to fitting, which improved the fits. The second is that the previous study relied upon polysomnographic data that was scored by visual pattern matching (i.e., expert hand-scoring); note, the autoscoring had 11% more N2 and 34% more Wake than the hand scored. Our analysis did not determine which of the two methods was ‘correct’, but the use of auto-scoring removes the significant inter-rater variability in sleep scoring and yields more uniform results (22). It is possible, therefore, that are more consistent scoring of N2 allowed for the identification of this stage as important to self-reported sleep depth and restfulness.

One curious finding is that we observe a slight increase in sleep satisfaction with older age, rather than the typically reported decline in sleep quality with increasing age (38–44). The increase we observe, however, is both small and not monotonically associated with the increase in age. Some studies suggest that when overall physical health factors are considered, a decline in sleep satisfaction is not an inevitable consequence of aging (43). The population we examined was predominantly healthy (Table 1), making the lack of decline in sleep satisfaction not altogether unexpected. Further, we want to emphasize that the age range of the population studied here was relatively narrow (62.7 ± 11.3 years) and may not extrapolated to changes over the lifespan.

While our data indicated that nearly a third of the variation in self-reported sleep quality could be predicted with the included parameters, there are other parameters that could have been derived and may offer additional explanatory power (e.g., variation in power spectral content across the night). Furthermore, our data set consists of only a single night of polysomnography and accompanying sleep quality data. Repeated measures from the same individuals could have provided greater clarity in terms of whether different subsets of individuals consistently responded to different aspects of sleep. Increasing the breadth of questions about the subjective experience of sleep may have also provided more insight, though there are a limited number of validated questionnaires that assess short-term sleep quality perception (45). Future studies examining orthogonal constructs of the subjective sleep experience would aid in a better understanding of the contributions of objective measures of sleep and their relationship to this subjective experience. Future studies examining orthogonal constructs of the subjective sleep experience would aid in a better understanding of the contributions of objective measures of sleep and their relationship to this subjective experience. The cohort was mainly middle-aged and older adults who were White and married. Thus, the physiologic underpinnings of subjective sleep quality in children or young adults, or individuals from other ethnic backgrounds was not addressed. Additionally, a relatively large subset of the data had to be excluded due to insufficient data quality or quantity, though this appeared to occur randomly and without specific demographic or medical bias.

It will be important for future experiments to determine whether active manipulation of the specific aspects of sleep identified here (e.g., increasing N2 at the expense of wake) results in an improvement in self-reported sleep quality. It will be important as well to determine whether changing N2 and wake are sufficient to improve clinical symptomatology in patient populations (e.g., insomnia). For example, we found that modest decreases in WASO (30 minutes) and increases in sleep efficiency (7-8%) were associated with higher ratings of sleep quality. These improvements are consistent with meta-analyses showing improvements of similar magnitude following Cognitive Behavioral Therapy for Insomnia (46). As such, targeted behavioral, pharmacological, or device-based manipulation of sleep could yield improved self-reported sleep quality and address the question of whether sleep quality begets quality sleep.

Acknowledgements.

The automated computation of sleep stages was performed on the Sherlock cluster. We would like to thank Stanford University and the Stanford Research Computing Center for providing computational resources and support that contributed to these research results. The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University). The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute (R24 HL114473, 75N92019R002)]. This research was supported by the Mental Illness, Research, Education, and Clinical Center (Department of Veterans Affairs, Palo Alto Division).

References.

1. Elder SJ, Pisoni RL, Akizawa T, Fissell R, Andreucci VE, Fukuhara S, et al. (2008): Sleep quality predicts quality of life and mortality risk in hemodialysis patients: Results from the Dialysis

Outcomes and Practice Patterns Study (DOPPS). Nephrology Dialysis Transplantation 23: 998–1004.

2. Jennings JR, Muldoon MF, Hall M, Buysse DJ, Manuck SB (2007): Self-reported sleep quality is associated with the metabolic syndrome. Sleep 30: 219–223. [PubMed: 17326548]
3. Tsai YW, Kann NH, Tung TH, Chao YJ, Lin CJ, Chang KC, et al. (2012): Impact of subjective sleep quality on glycemic control in type 2 diabetes mellitus. Family Practice 29: 30–35. [PubMed: 21795758]
4. Fiorentini A, Valente R, Perciaccante A, Tubani L (2007): Sleep's quality disorders in patients with hypertension and type 2 diabetes mellitus. International Journal of Cardiology 114: 50–52. [PubMed: 16675046]
5. Lao XQ, Liu X, Deng HB, Chan TC, Ho KF, Wang F, et al. (2018): Sleep quality, sleep duration, and the risk of coronary heart disease: A prospective cohort study with 60, 586 adults. Journal of Clinical Sleep Medicine 14: 109–117. [PubMed: 29198294]
6. Ritsner M, Kurs R, Ponizovsky A, Hadjez J (2004): Perceived quality of life in schizophrenia: Relationships to sleep quality. Quality of Life Research 13: 783–791. [PubMed: 15129888]
7. Jovevska S, Richdale AL, Lawson LP, Uljarević M, Arnold SRC, Trollor JN (2020): Sleep quality in autism from adolescence to old age. Autism in Adulthood 2: 152–162.
8. Ağargün Y, Hayrettin K, Solmaz M (1997): Subjective sleep quality and suicidality in patients with major depression. Journal of Psychiatric Research 31: 377–381. [PubMed: 9306295]
9. O'Donnell D, Silva EJ, Münch M, Ronda JM, Wang W, Duffy JF (2009): Comparison of subjective and objective assessments of sleep in healthy older subjects without sleep complaints. Journal of Sleep Research 18: 254–263. [PubMed: 19645969]
10. Baekeland F, Hoy P (1971): Reported vs recorded sleep characteristics. Archives of General Psychiatry 24: 548–551. [PubMed: 4325303]
11. Westerlund A, Lagerros YT, Kecklund GG, Axelsson J, Åkerstedt T (2016): Relationships between questionnaire ratings of sleep quality and polysomnography in healthy adults. Behavioral sleep medicine 14: 185–199. [PubMed: 25384098]
12. Hoch CC, Reynolds CF, Kupfer DJ, Berman SR, Houck PR, Stack JA (1987): Empirical note: self-report versus recorded sleep in healthy seniors. Psychophysiology 24: 293–299. [PubMed: 3602285]
13. Riedel BW, Lichstein KL (1998): Objective sleep measures and subjective sleep satisfaction: How do older adults with insomnia define a good night's sleep? Psychology and Aging 13: 159–163. [PubMed: 9533198]
14. Bonnet MH, Johnson LC (1978): Relationship of arousal threshold to sleep stage distribution and subjective estimates of depth and quality of sleep. Sleep 1: 161–168. [PubMed: 227030]
15. Akerstedt T, Hume K, Minors D, Waterhouse J (1994): The meaning of good sleep: a longitudinal study of polysomnography and subjective sleep quality. Journal of Sleep Research 3: 152–158. [PubMed: 10607120]
16. Laffan A, Caffo B, Swihart BJ, Punjabi NM (2010): Utility of sleep stage transitions in assessing sleep continuity. Sleep 33: 1681–1686. [PubMed: 21120130]
17. Kaplan KA, Hirshman J, Hernandez B, Stefanick ML, Hoffman AR, Redline S, et al. (2017): When a gold standard isn't so golden: Lack of prediction of subjective sleep quality from sleep polysomnography. Biological Psychology 123: 37–46. [PubMed: 27889439]
18. Kaplan KA, Hardas PP, Redline S, Zeitzer JM (2017): Correlates of sleep quality in midlife and beyond: a machine learning analysis. Sleep Med 34: 162–167. [PubMed: 28522086]
19. Faerman A, Kaplan KA, Zeitzer JM (2020): Subjective sleep quality is poorly associated with actigraphy and heart rate measures in community-dwelling older men. Sleep Medicine 73: 154–161. [PubMed: 32836083]
20. Hobson JA (1969): A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Electroencephalography Clinical Neurophysiology 26: 644.
21. Berry R, Brooks R, Gamaldo CE, Harding S, Lloyd R, Marcus C, Vaughn B (2015): The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. American Academy of Sleep Medicine, Version 2. Darien, Illinois.

22. Ferri R, Ferri P, Colognola RM, Petrella MA, Musumeci SA, Bergonzi P (1989): Comparison between the results of an automatic and a visual scoring of sleep EEG recordings. Sleep 12: 354–362. [PubMed: 2762689]
23. Rosenberg RS, Van Hout S (2013): The American Academy of Sleep Medicine inter-scorer reliability program: Sleep Stage Scoring. Journal of Clinical Sleep Medicine 9: 81–87. [PubMed: 23319910]
24. Stephansen JB, Olesen AN, Olsen M, Ambati A, Leary EB, Moore HE, et al. (2018): Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. Nature Communications 9: 1–15.
25. Cesari M, Stefani A, Penzel T, Ibrahim A, Hackner H, Heidbreder A, et al. (2021): Interrater sleep stage scoring reliability between manual scoring from two European sleep centers and automatic scoring performed by the artificial intelligence–based Stanford–STAGES algorithm. Journal of Clinical Sleep Medicine 17.
26. Younes M, Raneri J, Hanly P (2016): Staging sleep in polysomnograms: Analysis of inter-scorer variability. Journal of Clinical Sleep Medicine 12: 885–894. [PubMed: 27070243]
27. Quan SF, Howard BV, Iber C, Kiley JP, Nieto FJ, O'Connor GT, et al. (1997): The Sleep Heart Health Study: Design, rationale, and methods. Sleep 20: 1077–1085. [PubMed: 9493915]
28. Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, et al. (2018): The National Sleep Research Resource: Towards a sleep data commons. Journal of the American Medical Informatics Association 25: 1351–1358. [PubMed: 29860441]
29. Rechtschaffen A, Kales A (1968): A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. Los Angeles: Brain Info Service and Brain Res Inst.
30. Cohen J (1988): Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates.
31. Johns MW (1991): A new method for measuring daytime sleepiness: The Epworth sleepiness scale. Sleep 14: 540–545. [PubMed: 1798888]
32. Ware J, Sherbourn CD (1992): The MOS 36-Item short-form health survey (SF-36): I. Conceptual framework and item. Medical Care 30: 473–483. [PubMed: 1593914]
33. Honaker J, King G, Blackwell M (2011): Amelia II: A program for missing data. Journal of Statistical Software 45: 1–47.
34. Breiman L (2001): Random forests. Machine Learning 45: 5–32.
35. Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Gräler B (2018): Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 2018: e5518.
36. Borbely AA (1982): A two process model of sleep regulation. Human Neurobiology 1: 195–204. [PubMed: 7185792]
37. Dijk DJ, Czeisler CA (1995): Contribution of the circadian pacemaker and the sleep homeostat to sleep propensity, sleep structure, electroencephalographic slow waves, and sleep spindle activity in humans. The Journal of neuroscience: the official journal of the Society for Neuroscience 15: 3526–3538. [PubMed: 7751928]
38. Ancoli-Israel S, Ayalon L, Salzman C (2008): Sleep in the elderly: Normal variations and common sleep disorders. Harvard Review of Psychiatry, 16:279–286. [PubMed: 18803103]
39. Ancoli-Israel S, Kripke DF (1991): Prevalent sleep problems in the aged. Biofeedback and Self-Regulation 16: 349–359. [PubMed: 1760457]
40. Ancoli-Israel S (2009): Sleep and its disorders in aging populations. Sleep Medicine 10: 7–11. [PubMed: 18482864]
41. Bliwise DL (1993): Sleep in normal aging and dementia. Sleep, 16: 40–81. [PubMed: 8456235]
42. Scullin MK, Bliwise DL (2015): Sleep, cognition, and normal aging: Integrating a half century of multidisciplinary research. Perspectives on Psychological Science 10: 97–137. [PubMed: 25620997]
43. Bliwise DL, King AC, Harris RB, Haskell WL (1992): Prevalence of self-reported poor sleep in a healthy population aged 50-65. Social Science and Medicine 34: 49–55. [PubMed: 1738856]

44. Espiritu JRD (2008): Aging-related sleep changes. Clinics in Geriatric Medicine, 24: 1–14. [PubMed: 18035227]

45. Leppämäki S, Meesters Y, Haukka J, Lönnqvist J, Partonen T (2003): Effect of simulated dawn on quality of sleep - a community-based trial. BMC Psychiatry 5: 1–5.

46. Trauer JM, Qian MY, Doyle JS, Rajaratnam SMW, Cunnington D (2015): Cognitive Behavioral Therapy for chronic insomnia: A systematic review and meta-analysis. Annals of Internal medicine 163: 191–204. [PubMed: 26054060]

Highlights

- Polysomnographic sleep variables can explain 30% of the variance in subjective sleep satisfaction in middle and older aged adults.
- Sleep at the end of night is less informative about subjective sleep satisfaction than whole-night data.
- A decline in wakefulness coupled with a concomitant increase in N2 underlies a significant portion of our subjective sleep satisfaction.

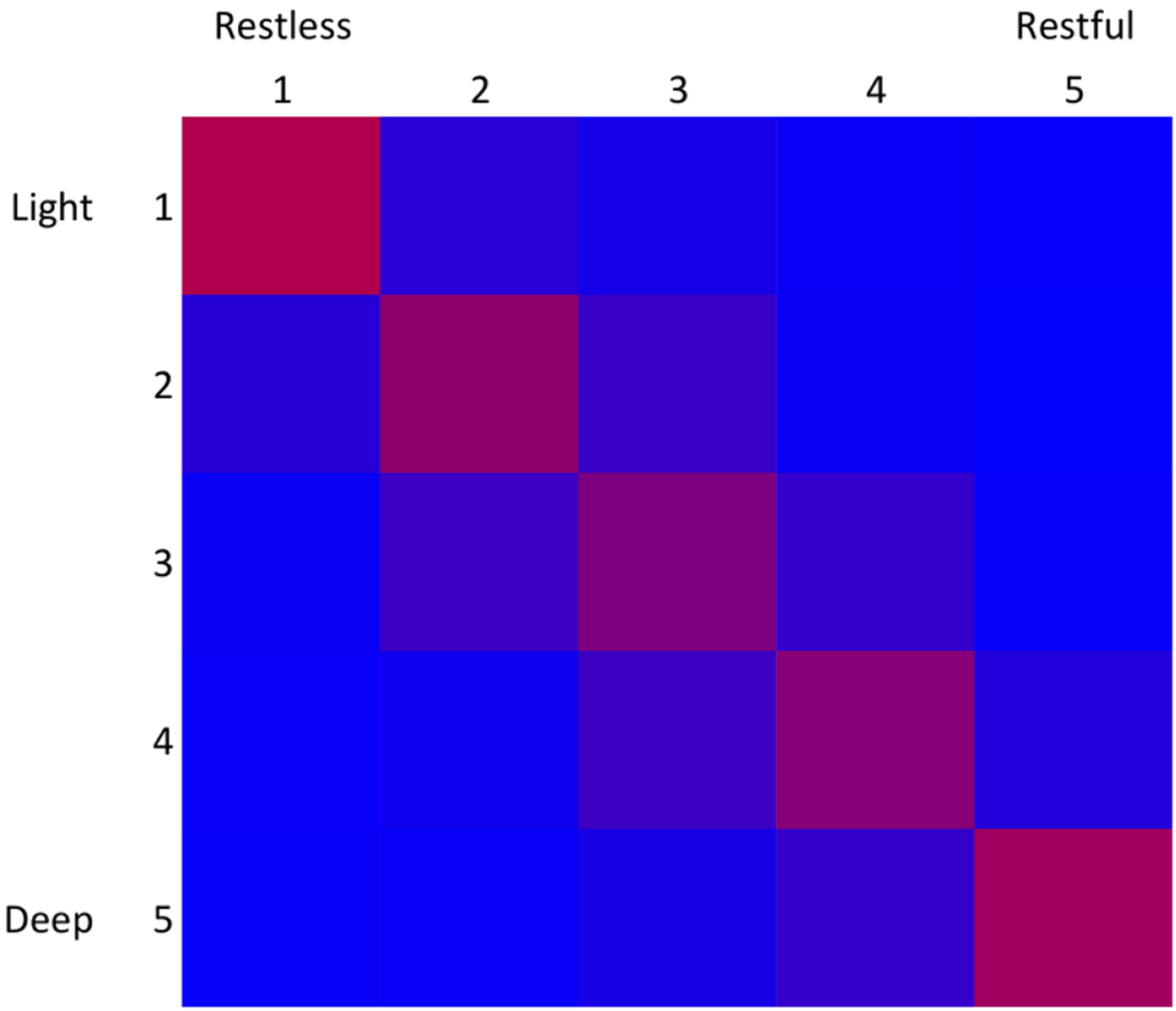

Figure 1.

Comparison of two self-reported sleep quality questions. Data from two questions are plotted as a heat map of the percent of time when a value in light/dark matched a value in restless/restful. Data are color coded with red (100%) and blue (0%) gradation.

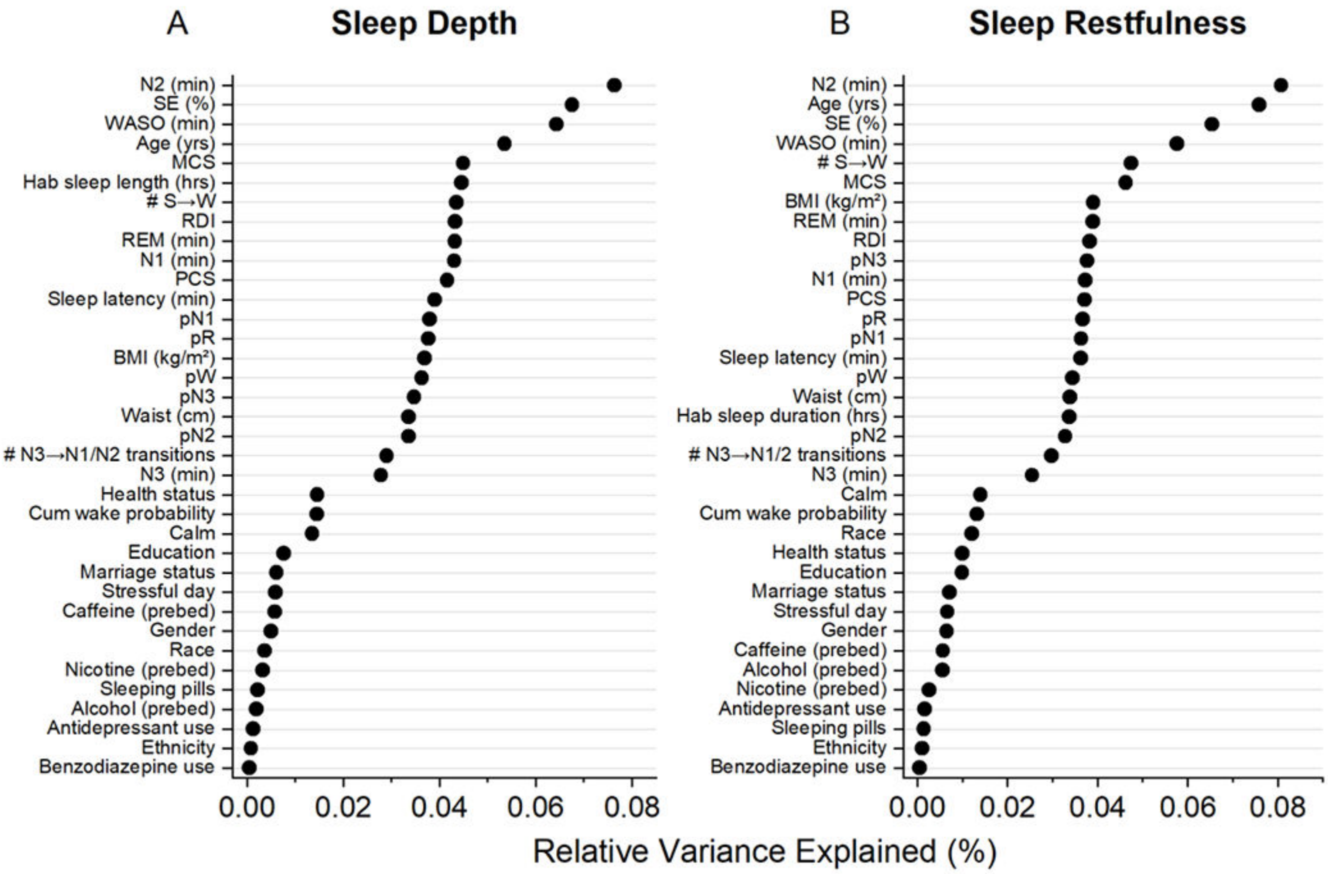

Figure 2.

Random Forest output for sleep depth (A) and restfulness (B). Data for individual predictors are plotted as variance explained relative to the total variance explained by the model (29.5% for depth, 26.8% for restfulness).

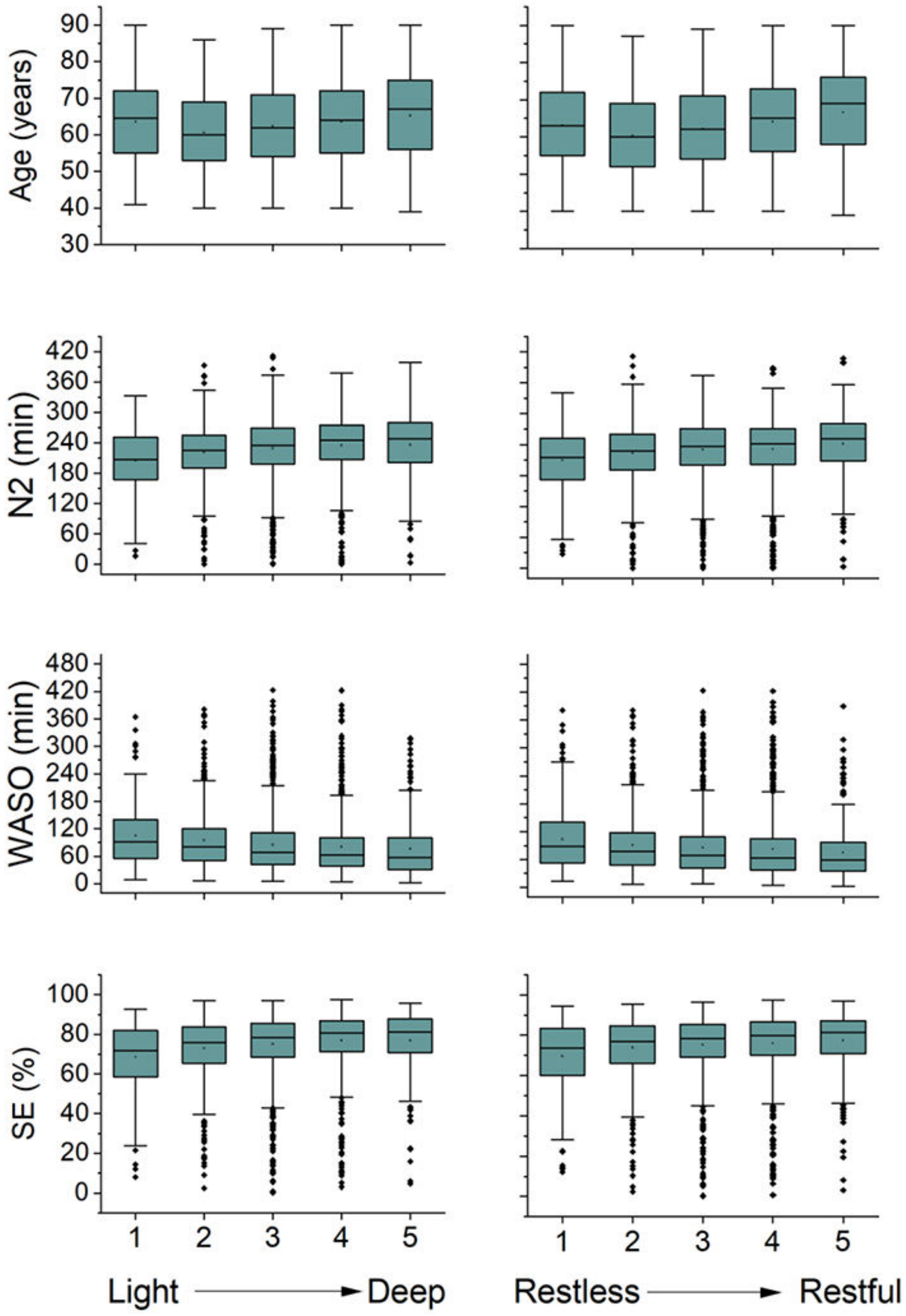

Figure 3.

Boxplots of age, minutes of N2, minutes of WASO, and SE for both sleep depth (left) and sleep restfulness (right). Progressive changes in each of these four variables is evident. Abbreviations: WASO (wake after sleep onset), SE (sleep efficiency).

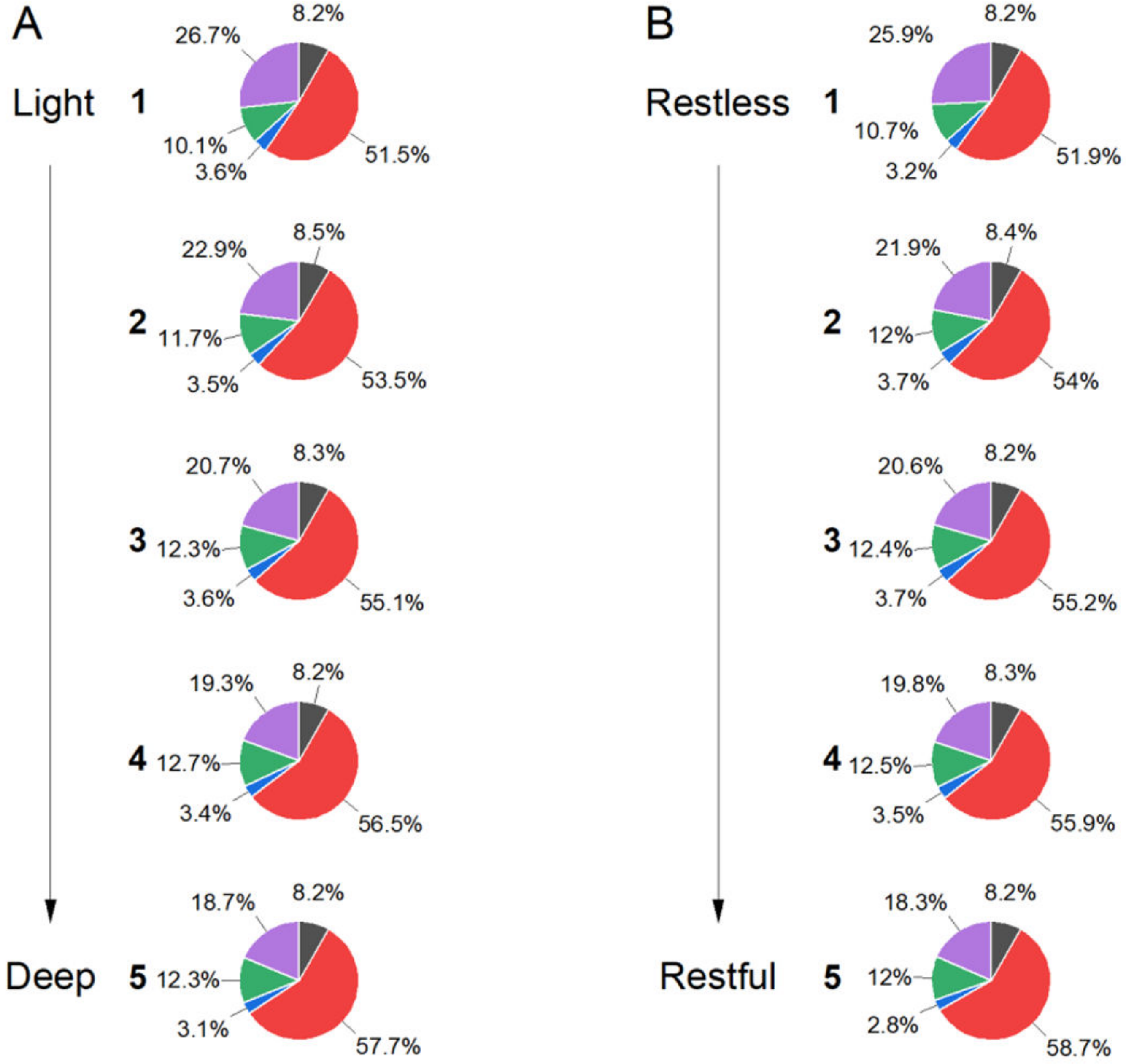

Figure 4.

Relative percent of each sleep stage by self-reported depth (A) or restfulness (B). Percent of time in bed after sleep onset spent in N1 (black), N2 (red), N3 (blue), REM (green), and wake (magenta) is shown. As the self-reported quality of sleep improves, progressively greater proportions of the night are spent in N2 with concomitant reductions in wake. No consistent changes are observed in N1, N3, or REM. Data represent the total number of epochs for each group.

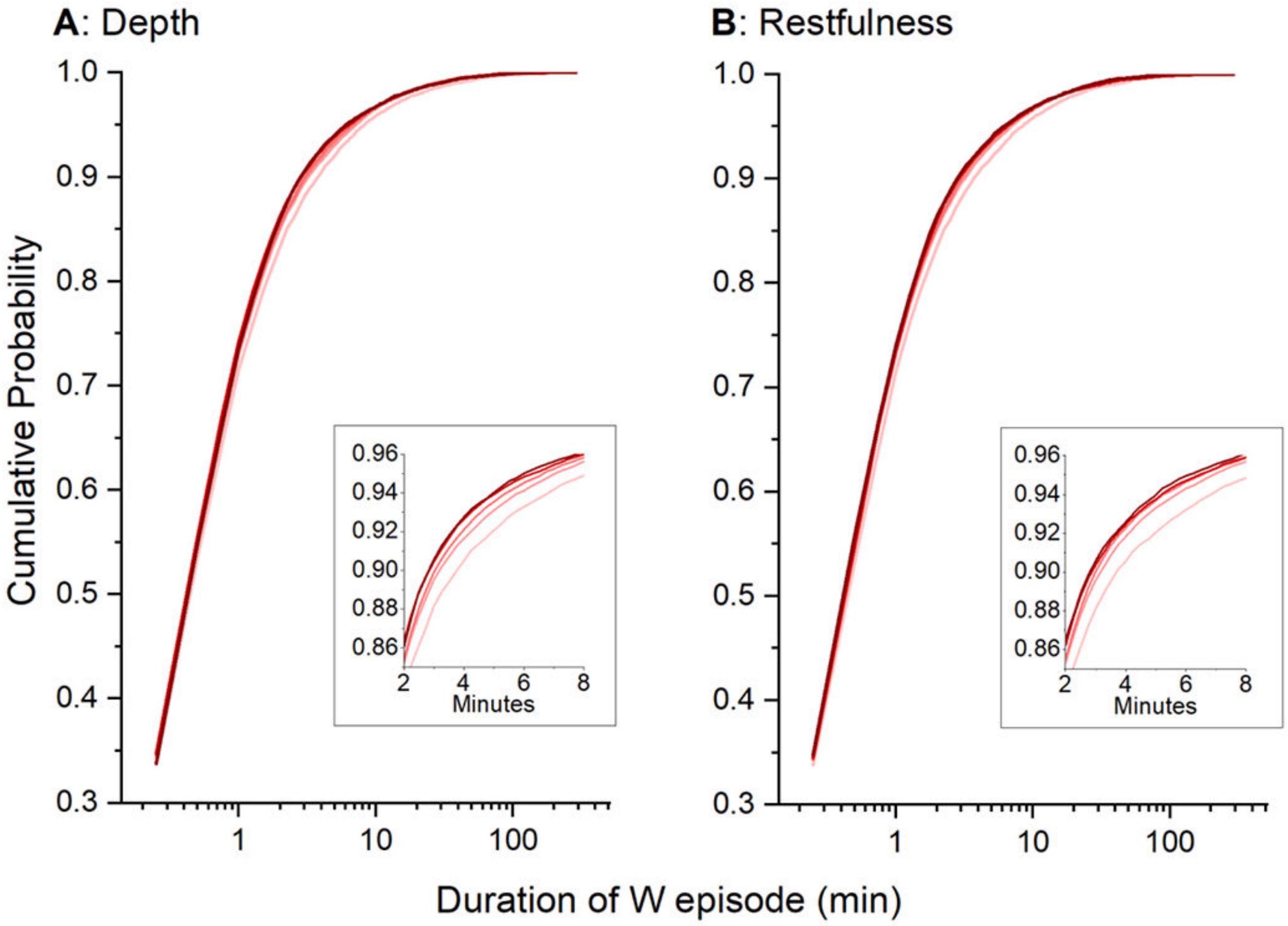

Figure 5.

The cumulative probability of wake epsiode lengths are plotted for self-reported sleep depth (A) and restfulness (B). Scores on these scales are color coded such that the darker red = 5 and progressively pinker colors are lower.

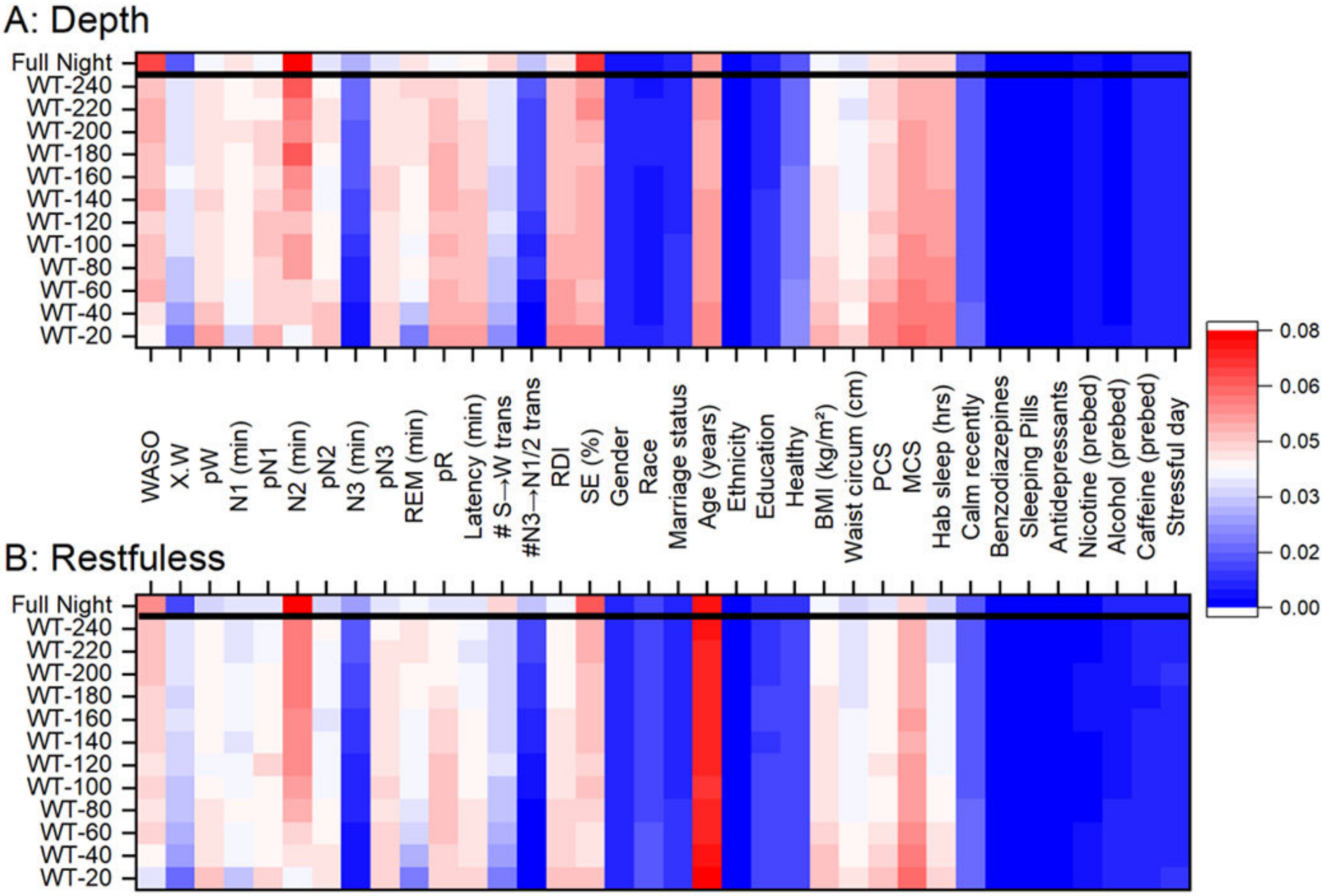

Figure 6.

A heat map of the relative amount of variance explained by different predictor variables for self-reported sleep depth (A) and restedness (B) in independent Random Forest models. Percent of variance explained is color coded as indicated. Each horizontal line represents a different amount of data from the full night of data to sequential parsing of the data starting at wake time (WT) (e.g., WT-120 is based on data between wake time and 120 minutes before wake time). Abbreviations: WASO: wake after sleep onset; X.W: average probability of wake during scored wake epochs multipled by the total duration of wake epochs; pW: average probability of wake; pN1: average probability of N1; pN2: average probability of N2; pN3: average probability of N3; REM: rapid eye movement sleep; pR: average probability of REM; Latency: sleep onset latency; #S→W trans: number of transitions between sequential epochs scored as sleep and then wake; N3→N1/2 trans: number of transitions between sequential epochs scored as N3 and then either N1 or N2 sleep; RDI: respiratory distress index; SE: sleep efficiency; BMI: body mass index; PCS: Physical Component Scale; MCS: Mental Component Scale; Hab sleep: habitual sleep length.

Table 1:

Baseline characteristics of the subset of the Sleep Heart Health Study population. Values are shown as mean ± SD, unless categorized by number and percentage. Abbreviations: SF-36 (36 item short form health survey), REM (rapid eye movement), WASO (wake after sleep onset), X probability (average matching probability of a state when the epoch is scored as that state). All parameters were included in random forest analysis.

Parameter	Response	Number
Gender	Female	1638 (52%)
	Male	1527 (48%)
Race	White	2694 (85%)
	Black	234 (7%)
	Other	237 (7%)
Ethnicity	Latinx	176 (6%)
Age (years)		62.7 ± 11.3
Marital status	Married	2485 (79%)
	Widowed	257 (8%)
	Divorced/Separate	313 (10%)
	Never Married	98 (3%)
	Unknown	12 (0%)
Education level	<10 years	241 (8%)
	11-15 years	1622 (51%)
	16-20 years	1163 (37%)
	>20 years	139 (4%)
Waist circumference (cm)		96.2 ± 13.6
Body Mass Index (BMI, kg/m^2)		27.9 ± 4.98
Physical Component Scale (PCS, SF-36)		47.9 ± 9.58
Mental Component Scale (MCS, SF-36)		53.3 ± 8.18
My health is excellent	Definitely true	685 (22%)
	Mostly true	1620 (51%)
	Not sure	394 (12%)
	Mostly false	283 (9%)
	Definitely false	183 (6%)
During the past 4 weeks, how much of the time have you felt calm and peaceful?	All of the time	267 (8%)
	Most of the time	1404 (44%)
	A good bit of the time	620 (20%)
	Some of the time	550 (17%)
	A little of the time	238 (8%)

Parameter	Response	Number
	None of the time	86 (3%)
Regular use of sleeping pills?	Yes	192 (6%)
Regular use of antidepressants?	Yes	218 (7%)
Use of benzodiazepines within 2 weeks of PSG?	Yes	182 (6%)
Epworth Sleepiness Scale		7.70 ± 4.34
Habitual sleep duration (hours)		7.13 ± 1.15
How stressful a day today?	A typical day Less stressful than usual More stressful than usual	2122 (67%) 568 (18%) 475 (15%)
Nicotine before bed (# products)		0.359 ± 1.36
Alcohol before bed (# drinks)		0.217 ± 0.689
Caffeine before bed (# drinks)		0.280 ± 0.654
N1 (minutes)		34.3 ± 22.5
N1 probability		53.1%
N1 (%)		8.3% ± 5.2%
N2 (minutes)		227. ± 61.3
N2 probability		84.0%
N2 (%)		55% ± 13%
N3 (minutes)		14.3 ± 21.9
N3 probability		64.9%
N3 (%)		3.5% ± 5.4%
REM (minutes)		50.1 ± 36.3
REM probability		75.4%
REM (%)		12% ± 8.5%
WASO (minutes)		86.8 ± 63.1
W probability		84.1%
WASO (%)		21% ± 15%
Sleep → Wake shifts (#)		6.5 ± 3.1
N3 → N1 or N2 shifts (#)		1.2 ± 1.4

Parameter	Response	Number
Sleep latency (minutes)		24.1 ± 23.2
Sleep efficiency (%)		74.7 ± 15.4
Respiratory Disturbance Index		8.26 ± 12.0
Self-reported sleep quality: Light (1) vs. Deep (5)	1	272 (9%)
	2	572 (18%)
	3	1290 (41%)
	4	740 (23%)
	5	291 (9%)
Self-reported sleep quality: Restless (1) vs. Restful (5)	1	361 (11%)
	2	727 (23%)
	3	997 (32%)
	4	749 (24%)
	5	331 (10%)