Received: 24 January 2021 | Revised: 30 March 2021 | Accepted: 20 April 2021

DOI: 10.1111/jsr.13386

RESEARCH ARTICLE

Physiological sleep measures predict time to 15-year mortality in community adults: Application of a novel machine learning framework

Meredith L. Wallace[1,2] | Timothy S. Coleman[2] | Lucas K. Mentch[2] | Daniel J. Buysse[1] | Jessica L. Graves[3] | Erika W. Hagen[4] | Martica H. Hall[1] | Katie L. Stone[5] | Susan Redline[6] | Paul E. Peppard[4]

[1]Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA

[2]Department of Statistics, University of Pittsburgh, Pittsburgh, PA, USA

[3]University of Pittsburgh Medical Center, Pittsburgh, PA, USA

[4]Department of Population Health Sciences, University of Wisconsin, Madison, WI, USA

[5]California Pacific Medical Center Research Institute, San Francisco, CA, USA

[6]Departments of Medicine, Brigham and Women's Hospital, Harvard Medical School, Beth Israel Deaconess Medical Center, Boston, MA, USA

Correspondence

Meredith L. Wallace, Department of Psychiatry, University of Pittsburgh, 3811 O'Hara Street 15213, Pittsburgh, PA, USA.
Email: lotzmj@upmc.edu

Funding information

National Institute on Aging, Grant/Award Number: R01AG036838, R01AG047139, R01AG056331 and R01AG14124; National Heart, Lung, and Blood Institute, Grant/Award Number: R01HL62252, U01HL53916, U01HL53931, U01HL53934, U01HL53937, U01HL53938, U01HL53940, U01HL53941, U01HL64360, R01AG047139 and HL114473; National Institute of Aging, Grant/Award Number: 1R01AG036838 and R01AG14124; National Center for Research Resources, Grant/Award Number: 1UL1RR025011

Summary

Clarifying whether physiological sleep measures predict mortality could inform risk screening; however, such investigations should account for complex and potentially non-linear relationships among health risk factors. We aimed to establish the predictive utility of polysomnography (PSG)-assessed sleep measures for mortality using a novel permutation random forest (PRF) machine learning framework. Data collected from the years 1995 to present are from the Sleep Heart Health Study (SHHS; *n* = 5,734) and the Wisconsin Sleep Cohort Study (WSCS; *n* = 1,015), and include initial assessments of sleep and health, and up to 15 years of follow-up for all-cause mortality. We applied PRF models to quantify the predictive abilities of 24 measures grouped into five domains: PSG-assessed sleep (four measures), self-reported sleep (three), health (eight), health behaviours (four), and sociodemographic factors (five). A 10-fold repeated internal validation (WSCS and SHHS combined) and external validation (training in SHHS; testing in WSCS) were used to compute unbiased variable importance metrics and associated *p* values. We observed that health, sociodemographic factors, and PSG-assessed sleep domains predicted mortality using both external validation and repeated internal validation. The PSG-assessed sleep efficiency and the percentage of sleep time with oxygen saturation <90% were among the most predictive individual measures. Multivariable Cox regression also revealed the PSG-assessed sleep domain to be predictive, with very low sleep efficiency and high hypoxaemia conferring the highest risk. These findings, coupled with the emergence of new low-burden technologies for objectively assessing sleep and overnight oxygen saturation, suggest that consideration of physiological sleep measures may improve risk screening.

KEYWORDS

hypoxaemia, rapid eye movement, risk screening, sleep efficiency

Redline and Peppard joint senior authors.

1 | INTRODUCTION

Healthcare practitioners routinely screen patients for risk factors. However, despite spending roughly a third of our lives engaged in sleep and disturbed sleep's association with adverse outcomes (Buysse, 2014), sleep is absent from high-profile health screening recommendations (Swenson & Ebell, 2016). Clarifying the predictive utility of sleep for all-cause mortality, a health outcome of ubiquitous importance, relative to established health risk factors could inform screening recommendations and guide treatment and mechanistic research.

Physiological measures such as those derived from polysomnography (PSG) are particularly important for characterising between-individual differences in sleep. Considered the "gold standard" for sleep measurement, PSG uses electrophysiological measures to quantify sleep and wake states and time in specific sleep stages. PSG also captures measures of abnormal systemic physiology during sleep, e.g. sleep-disordered breathing (SDB). SDB is prevalent (Senaratna et al., 2017), negatively affects sleep quality and architecture (Redline et al., 2004), and is a known risk factor for mortality (Punjabi et al., 2009; Young et al., 2008).

Despite evidence of the predictive abilities of various individual PSG measures for mortality, it is still unclear whether they are useful above and beyond a robust set of established health risk factors and, if so, which specific PSG measures have the strongest associations. A major obstacle to making this determination is that sleep and non-sleep risk factors likely have complex and non-linear relationships with each other, especially with regards to their combined downstream effects on health and mortality. However, standard regression approaches tend to over-simplify these underlying relationships. Any assessment of the predictive utility of PSG-assessed sleep measures for mortality should acknowledge this underlying complexity, and model it accordingly.

Random forests are a flexible, powerful, and top-performing machine learning tool for modelling complex relationships (Fernández-Delgado, Cernadas, Barro, & Amorim, 2014). As we have demonstrated previously for self-report and actiography-assessed sleep (Wallace, Buysse, et al., 2019; Wallace, Lee, et al., 2019), this technique may yield new information regarding the predictive abilities of PSG-assessed sleep measures that cannot be observed with traditional regression models. However, commonly used approaches to extract the predictive importance of variables within random forests suffer from bias towards predictors that are even moderately correlated, which poses a major barrier to model interpretation (Nicodemus, Malley, Strobl, & Ziegler, 2010; Strobl, Boulesteix, Zeileis, & Hothorn, 2007). Therefore, we introduce and apply for the first time a novel permutation random forest (PRF) machine-learning framework that produces *unbiased* rankings of variable importance and which facilitates formal hypothesis testing of whether a variable, or group of variables, predicts mortality (Coleman, Peng, & Mentch, 2019; Mentch & Hooker, 2017).

In the present study, we apply PRF machine learning to harmonised data (total $n = 6{,}749$) from the Sleep Heart Health Study (SHHS; $n = 5{,}734$) (Quan et al., 1997) and the Wisconsin Sleep Cohort Study (WSCS; $n = 1{,}015$) (Young et al., 2008). Both studies conducted PSG in well-characterised community-based cohorts of middle-aged and older adults and have at least 15 years of follow-up for all-cause mortality. Our primary aim was to establish the predictive utility of a group of PSG measures for mortality relative to other established groups of risk factors (health, health behaviours, and sociodemographic factors). Given our interest in informing health screening recommendations, we prioritised more scalable PSG measures. We also examined whether the combination of both habitual self-reported sleep and PSG measures would enhance prediction and explored which individual sleep characteristics are most predictive.

2 | METHODS

2.1 | Study designs, participants, and procedures for parent cohorts

The SHHS is a multicentre cohort study to determine the cardiovascular and other health consequences of SDB in community adults, with the initial examination cycle occurring between 1995 and 1998. Participants were aged ≥40 years at the initial PSG visit and had no current treatment of sleep apnoea, tracheostomy, or nocturnal oxygen therapy. We utilised an open-source dataset of $n = 5{,}804$ SHHS participants available from the National Sleep Research Resource (www.sleepdata.org; Dean et al., 2016; Zhang et al., 2018).

The WSCS is a longitudinal cohort study to investigate the course of SDB and other sleep disorders. Starting in 1988, a random sample of community adults aged 30–60 years were recruited from payroll records of Wisconsin state agencies and followed at ~4-year intervals. We utilised a dataset of $n = 1{,}184$ individuals for whom specific data-sharing consent was obtained and in whom digitally recorded PSG was captured after the year 2000.

Participants in the SHHS underwent Type 2 PSG at home using an 11-channel portable monitor (Compumedics P-series, Abbotsford) at each of two examination cycles. Participants in the WSCS underwent 1 night of laboratory-based video-assisted PSG at each visit using either a 16-channel (years 2000–2009) or 19-channel (years 2009–2015) system (Grass Instruments). In both studies, PSG scoring was conducted by trained research technicians in accredited sleep laboratories using established quality control procedures, and was blinded to other data. Research technicians also performed clinical interviews and collected objective health measures. Additional study details related to study designs are provided in the Supplement (Sections 1–2 in Appendix S1) and prior publications provide details of PSG scoring and quality control measures for each cohort (Hori et al., 2001; Peppard et al., 2013; Quan et al., 1997; Rechtschaffen & Kales, 1968; Redline et al., 1998; Whitney et al., 1998; Young et al., 1993).

All participants provided informed consent. This study was approved by the University of Pittsburgh Institutional Review Board (PRO17050218).

TABLE 1 Descriptive statistics for the combined (WSCS + SHHS), WSCS, and SHHS cohorts

	Combined **(*n* = 6,749)**	**WSCS** **(*n* = 1,015)**	**SHHS** **(*n* = 5,734)**
Sociodemographic factors			
Age, years, mean (*SD*)	63.01 (10.83)	62.37 (8.64)	63.12 (11.17)
Female (versus male), *n* (%)	3,482 (51.59)	477 (47.00)	3,005 (52.41)
White (versus non-White), *n* (%)	5,818 (86.21)	955 (94.09)	4,863 (84.81)
Education ≥16 years (versus <16 years), *n* (%)	2,779 (41.18)	472 (46.5)	2,307 (40.23)
Married or living as married (versus other), *n* (%)	5,289 (78.37)	740 (72.91)	4,549 (79.33)
Health			
Total cholesterol, mg/dl, mean *(SD)*	205.17 (38.16)	191.67 (37.49)	207.56 (37.78)
Number of cardiovascular risk factors, mean (*SD*)	0.62 (0.92)	0.60 (0.86)	0.63 (0.93)
Diabetes, *n* (%)	527 (7.81)	134 (13.20)	393 (6.85)
Systolic blood pressure, mmHg, mean (*SD*)	127.48 (18.68)	128.35(15.38)	127.33(19.2)
Body mass index, kg/m^2, mean (*SD*)	28.54 (5.4)	30.58(6.6)	28.18(5.07)
SF-36 Physical Component Score, mean (*SD*)	47.85 (9.25)	49.37 (7.59)	47.58 (9.49)
SF-36 Mental Component Score, mean (SD)	53.45 (7.72)	54.05 (6.78)	53.34 (7.88)
Number of medication codes, mean (*SD*)	2.69 (2.57)	3.42 (2.59)	2.56 (2.54)
Health behaviours			
Cans of caffeinated soda/day, mean (*SD*)	0.51 (1.04)	0.64(1.04)	0.49(1.03)
Cups of coffee or tea/day, mean (*SD*)	2.14 (2.34)	1.99(1.8)	2.17(2.42)
No alcohol use, *n* (%)	3,367 (49.89)	246 (24.24)	3,121 (54.43)
Smoking status, *n* (%)			
Current smoker (versus past or non-smoker)	683 (10.12)	90 (8.87)	593 (10.34)
Past smoker (versus current or non-smoker)	2,868 (42.50)	409 (40.30)	2,459 (42.88)
Non-smoker (versus current or past)	3,198 (47.48)	516 (50.84)	2,682 (46.77)
Self-report sleep			
Frequency of difficulty getting back to sleep, mean (*SD*)	1.69 (1.03)	1.68 (1.01)	1.69 (1.03)
Frequency of excessive daytime sleepiness, mean (*SD*)	1.43 (0.97)	1.36 (1.02)	1.45 (0.96)
Sleep duration, hr, mean *(SD)*	7.17 (1.14)	7.27 (1.03)	7.15 (1.16)
PSG sleep			
Sleep efficiency, mean (*SD*)	82.18 (10.63)	78.65(10.71)	82.8(10.5)
% Total sleep time in Stage 3–4 (N3) sleep	16.78 (11.83)	8.08(7.85)	18.32(11.74)
% Total sleep time in rapid eye movement sleep	18.77 (6.92)	15.66(6.12)	19.33(6.9)
% Total sleep time spent with arterial oxygen saturation <90% (T90)	3.51 (10.47)	3.68(11.39)	3.48(10.3)
Outcome			
Time to mortality, median (range)	11.65 (1.01– 15.00)	7.77 (1.08– 15.00)	11.84 (1.01– 15.00)
All-cause mortality, *n* (%)	1,369 (20.28)	106 (10.44)	1,263 (22.03)

PSG, polysomnography; SF-36, 36-item Short Form Health Survey; SHHS, Sleep Heart Health Study; WSCS, Wisconsin Sleep Cohort Study.

2.2 | Analytic samples

For the present analysis, we developed analytical samples from the full SHHS and WSCS parent cohorts described above, utilising only one visit of data for each cohort. For the SHHS, we included data from the initial examination cycle. For the WSCS, we included data from each participant's last observed PSG visit, which provided more comparable age ranges across cohorts. To reduce the potential for confounding by disease processes, participants in both the SHHS and WSCS were selected if they had ≥1 year of follow-up for mortality. Participants in the WSCS were further selected if they had no use of sleep apnoea treatment and/or oxygen therapy to better harmonise the sample with the SHHS, which did not allow for current treatment or oxygen therapy. However, individuals with untreated and/or less severe sleep apnoea were allowed in both samples. As several variables had data assessed to be missing at random, we performed random forest multiple imputation (Stekhoven & Bühlmann, 2012) to retain the full analytical sample for analysis (Supplement Section 3 in Appendix S1). The final analytical samples were *n* = 5,734 for the SHHS and *n* = 1,015 for the WSCS. Table 1 provides characterisations of the analytical samples.

2.3 | Measures

The outcome was time to all-cause mortality, censored at 15 years to better equate follow-up windows across cohorts. In the SHHS, there was 22% mortality (1,263 deaths), with a median (25th–75th percentile, i.e. interquartile range [IQR]) follow-up of 11.84 (10.69–12.84) years. In the WSCS, there was 10% mortality (106 deaths) with a median (IQR) follow-up of 7.77 (5.91–11.66) years. See Supplement Section 4 for details in Appendix S1.

We initially harmonised 60 predictors, selected based on preliminary relevance and comparability across cohorts. As large numbers of the health and sleep predictors were highly correlated and overlapped conceptually, we used clinical/scientific expertise, guided by exploratory factor analysis, to select a relatively independent subset of predictors. When making selections, our end goals were: (a) maximising interpretability and relevance, (2) reducing type 1 error; (c) retaining a relatively balanced and equitable number of representative measures across domains, and (d) minimising impact of PSG scoring definitions and laboratory-specific scoring protocols. Details are provided in the Supplement Section 5 in Appendix S1. The final 24 predictors selected for primary analyses reflected five domains: sociodemographic factors (five predictors), health (eight predictors), health behaviours (four predictors), self-report sleep (three predictors), and PSG sleep (four predictors). They are listed below and described in Table 1.

2.3.1 | Sociodemographic factors

Age, education (≥16 versus <16 years), marital status (married or living as married versus other marital status), sex (female versus male), and race (White versus non-White).

2.3.2 | Health

Self-report of diabetes, number of cardiovascular risk factors (self-report of diagnosed heart attack, heart failure, hypertension, stroke, angina, angioplasty, coronary bypass, and pacemaker), the 36-item Short Form Health Survey (SF-36) Physical and Mental Component Standardised Scores (SF-36 PCS and MCS, respectively; Ware, Kosinski, & Gandek, 2000), measured body mass index (BMI), measured sitting systolic blood pressure, measured total cholesterol, and self-reported number of unique medication codes taken (see Supplement Section 6 for medication harmonisation details).

2.3.3 | Health behaviours

Smoking (none, past, current), alcohol use (any versus none), cans of caffeinated soda per day, and cups of coffee/tea per day. In both cohorts, participants were asked to report their usual caffeine intake per day, facilitating the use of more nuanced count measures. We used the categorisation of any versus none for alcohol use because the reporting time scales differed in the SHHS versus WSCS (per day versus per week), leading to uncertainty of the comparability of more nuanced alcohol frequency measures.

2.3.4 | Self-report sleep

Frequencies of excessive daytime sleepiness and difficulty with falling back to sleep (both rated 0 [never] to 4 [almost always]) and sleep duration (weighted average of self-reported usual hours of sleep on weekends and workdays).

2.3.5 | Polysomnography sleep

Sleep efficiency (SE; percentage of time in bed spent asleep), percentage of total sleep time in rapid eye movement (REM) sleep (REM %), percentage of total sleep time in Stages 3–4 non-REM sleep, corresponding with N3 or "slow-wave sleep" in current terminology (Stage 3%–4%), percentage of total sleep time in REM sleep (REM %), and the percentage of total sleep time spent with arterial oxygen saturation (SpO_2) <90% (T90), a measure of overnight hypoxaemia. For primary analyses, T90 was selected over the more common apnea–hypopnea index (AHI) because it is automatically derived, does not require scoring of arousals, is robust to laboratory and consensus definitions, and is easily scalable. As such it bolsters the possibility of developing highly generalisable risk prediction models. However, we also considered the AHI in post hoc exploratory analyses, as it is a conventional measure of sleep apnoea severity that reflects the intermittent nature of hypoxaemia and sleep fragmentation.

2.4 | Data analysis

Our primary modelling strategy was a novel PRF machine learning framework (Coleman et al., 2019). The PRF quantifies the predictive utility of a variable (or group of variables) using a permutation variable importance (pVIMP) metric. The pVIMP reflects the difference in misclassification error × 100 between a model including the variable under examination versus one including a randomly permuted version of that variable. Permuted variables should, by construction, hold no predictive power for the response. Thus, if the misclassification error of a model is meaningfully reduced when the original variable is used instead of its permuted counterpart, that variable is considered to provide unique predictive information. Unlike other importance metrics, the pVIMP is not biased towards correlated variables.

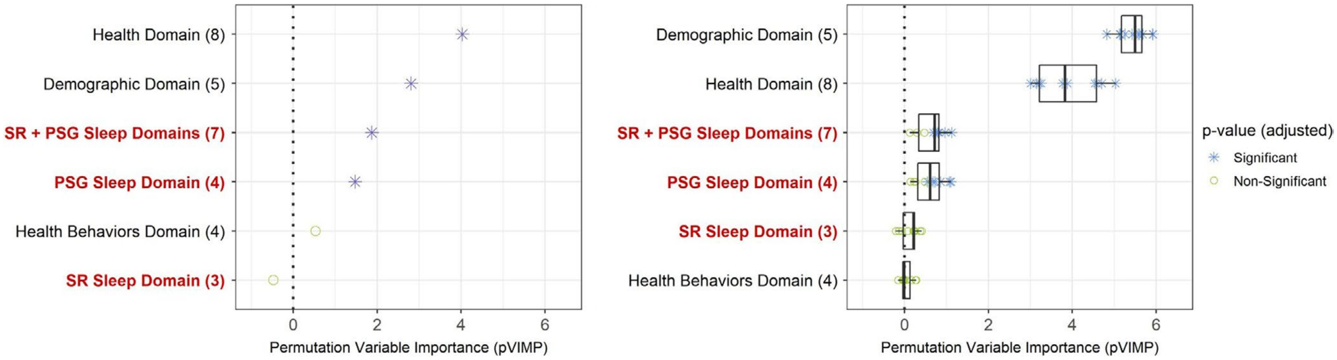

FIGURE 1 Importance of domains of predictors for all-cause mortality based on external validation (left panel) and internal validation (right panel). Significant pVIMPs have *p* values <.01. Red/bold text indicates sleep domains. PSG, polysomnography; SR, self-report

The pVIMP reflects change in misclassification error as measured by Harrell's C, equivalent to 1 area under the curve (AUC) (Heagerty & Zheng, 2005). We can interpret the size of the pVIMPs using existing AUC benchmarks. For example, the changes in AUC between a small to medium effect size and a medium to large effect size are ~0.08 (Rice & Harris, 2005). As such, inclusion of a predictor (or group of predictors) with a pVIMP of 8 would produce a change in model accuracy that could be considered highly perceptible (i.e., large enough to move between small to moderate or moderate to large AUC effect size). However, as effect size interpretations are ultimately context dependent, the sizes of the pVIMPs of the most well-established predictors for mortality (e.g. age, physical health) provide useful benchmarks for this application.

The hypothesis tests we employ to determine whether the misclassification error is meaningfully reduced (i.e. a test of whether pVIMP >0) utilise a permutation distribution of the null effect to directly indicate the proportion of the time that the observed pVIMP for a given predictor is larger than we observe by random chance (the *p* value). Because these tests utilise a permutation distribution rather than a sampling distribution, traditional confidence intervals (CIs) for the pVIMP are not readily available (Coleman et al., 2019; Mentch & Hooker, 2017).

We used both external validation and internal validation strategies to quantify and test the predictive utility of each domain of predictors, the combined PSG and self-report sleep domains, and each individual predictor. For external validation, we developed the model in the SHHS and then applied it to the WSCS to compute pVIMPs and *p* values. The SHHS was selected for training because it is well-powered for this analysis; the WSCS was selected for testing because, although it is not large enough for training, it is sufficiently powered for external validation (Collins, Ogundimu, & Altman, 2016). For internal validation, we used repeated mortality-stratified subsampling to develop 10 pairs of training/testing datasets based on the combined (SHHS + WSCS) sample. The pVIMPs and *p* values from the 10 testing samples were summarised using the median ($\text{pVIMP}_{\text{med}}$ and p_{med}).

We also fit a Cox proportional hazards regression model to all predictors simultaneously with the combined sample. We used categorial PSG variables to accommodate potentially non-linear relationships. We tested each domain of variables using a likelihood ratio test and explored effects of individual predictors using hazard ratios (HRs) and 95% CIs.

In exploratory analyses, we also considered the AHI. Using machine learning with internal and external validation, we tested whether the AHI was significant among the full set of predictors and whether its addition may improve the predictive accuracy of the full set of PSG measures. We also tested whether the AHI was significant if used in place of T90, as both the AHI and T90 represent sleep-related respiration. Finally, we tested whether the AHI was significant when added to the full Cox model.

To adjust for multiple comparisons in primary domain tests, we used a two-sided test of significance with $\alpha = 0.01$ (a conservative Bonferroni adjustment assuming five independent domains). We did not adjust for multiple comparisons in exploratory analyses per current recommendations (Cao & Zhang, 2014). Analyses were performed using R version 4.0.2. Supplement Sections 7–8 in Appendix S1 provide additional data analysis details.

3 | RESULTS

Participants in the SHHS and WSCS were roughly comparable on sociodemographic factors, with some differences in baseline profiles (e.g. higher BMI and alcohol use in the WSCS). Relative to the WSCS, SHHS participants had higher REM % and Stage 3%–4%, but lower SE (Table 1).

3.1 | Permutation random forest machine learning models

Using both internal and external validation strategies, the sociodemographic factors, health, combined PSG and self-reported sleep, and PSG sleep domains predicted time to all-cause mortality (Internal validation $pVIMP_{med}$ [p_{med}] = 5.50 [<.001], 3.82 [<.001], 0.72 [.001], 0.61 [.010]; External validation *pVIMP* [*p*] = 2.81 [<.001], 4.03 [<.001], 1.89 [<.001],1.47 [<.001], respectively). As the *pVIMPs* from the combined PSG and self-report sleep domain were only slightly higher than those from the PSG sleep domain alone, only an incremental improvement in prediction can be attributed to self-report sleep. The health behaviours and self-report sleep domains did not meaningfully contribute to the model (Figure 1).

Using external validation, age, PSG SE, the SF-36 PCS, T90, sex, and diabetes predicted time to all-cause mortality (*pVIMP[p]* = 1.79 [<.001], 0.76 [.010], 0.76 [.004], 0.72 [.008], 0.71 [.016], 0.70 [.026], respectively). Using internal validation, the same predictors were rated among the top six; however, only age, the SF-36 PCS and sex were predictive ($pVIMP_{med}[p_{med}]$ = 4.02 [<.001], 0.58 [.001], 0.33 [.039], respectively; Figure 2).

In exploratory analyses, the AHI was not significant in either external validation or internal validation, regardless of whether T90 was also included in the model. Including AHI among the set of PSG predictors increased the pVIMP by only a relatively small amount. See Supplement Section 9 for details.

3.2 | Cox proportional hazards model

All domains, except self-report sleep, predicted time to mortality (Table 2). SE, REM %, and T90 were significant individual PSG measures. Very low SE (quintile 1 [Q1]; SE <75%) increased mortality risk relative to moderate-to-high SE (Q3–4; SE of 82%–91%) with HRs of 1.27. However, very low SE did not differ from very high SE (Q5; SE >91%), suggesting a potentially non-linear relationship. The lowest REM % (Q1; REM % <13.5%) increased mortality risk relative to all other quintiles, with HRs between 1.26 and 1.39. Having the highest T90 (>75th percentile; T90 >1.85%) increased risk of mortality relative to those with low T90 (<50th percentile; T90 <0.20%). See Table 2 and Supplement Section 10 in Appendix S1 for all HRs (95% CIs).

In exploratory analyses, the AHI was not a significant predictor in the Cox model ($\chi^2 = 6.79$, $p = .148$), although SE, REM %, and T90 remained significant (SE $\chi^2 = 16.02$, $p = .003$; REM % $\chi^2 = 19.66$, $p = .0006$; T90 $\chi^2 = 12.97$, $p = .002$).

4 | DISCUSSION

Using a novel machine learning framework, a set of four PSG-assessed physiological measures (T90, SE, REM %, Stage 3%–4%) predicted time to all-cause mortality over 15 years of follow-up, above and beyond complex combinations of other established risk factors typically included in health screening recommendations. Although the set of PSG measures was less predictive than sets of traditional health and sociodemographic measures, it was more predictive than health behaviours and habitual self-reported sleep.

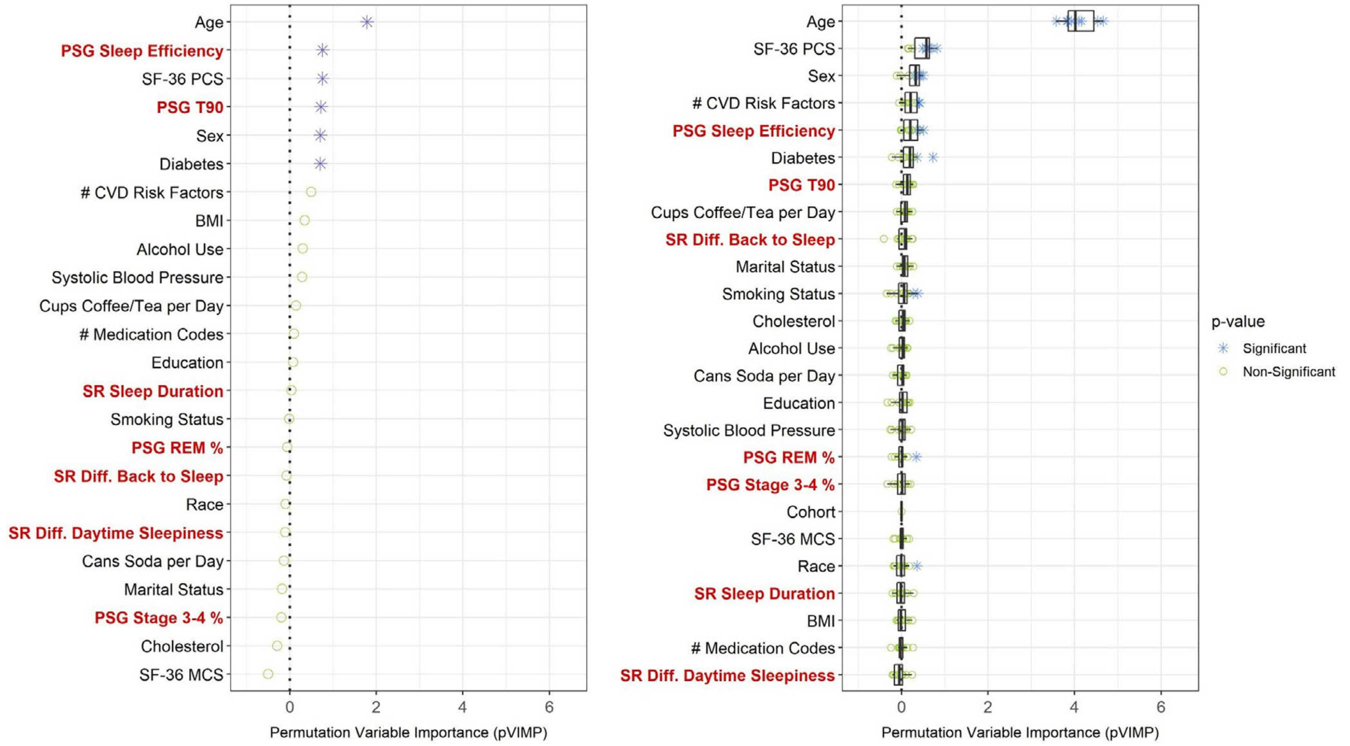

FIGURE 2 Importance of individual predictors for all-cause mortality based on external validation (left panel) and internal validation (right panel). Significant pVIMPs have *p* values <.05. Red/bold text indicates sleep predictors. BMI, body mass index; CVD, cardiovascular disease; PSG, polysomnography; REM %, percentage of total sleep time in rapid eye movement sleep; Stage 3%–4%, percentage of total sleep time in Stage 3–4 sleep; SF-36 PCS, 36-item Short Form Health Survey Physical Component Score; SF-36 MCS, SF-36 Mental Component Score; SR, self-report; T90, percentage of total sleep time spent with arterial oxygen saturation (SpO_2) <90%

TABLE 2 Multivariable Cox proportional hazards model in the combined (SHHS + WSCS) sample ($n = 6{,}749$)

Variable	HR (95% CI)	p
Sociodemographic factors domain: $\chi^2 = 791.55$, $p < .001$		
Age, standardised	2.72 (2.52–2.94)	<.001
Female (versus male)	0.67 (0.58–0.77)	<.001
White (versus non-White)	1.11 (0.93–1.32)	.240
Education ≥16 years (versus <16 years)	0.93 (0.82–1.05)	.255
Married or living as married (versus other marital status)	0.86 (0.76–0.98)	.027
WSCS (versus SHHS)	0.96 (0.77–1.21)	.753
Health domain: $\chi^2 = 460.07$, $p < .001$		
Total cholesterol, standardised	0.96 (0.91–1.02)	.185
Number of cardiovascular risk factors	1.17 (1.11–1.22)	<.001
Diabetes	1.48 (1.26–1.73)	<.001
Systolic blood pressure, standardised	1.07 (1.01–1.12)	.013
BMI: $\chi^2 = 29.54$, $p < .001$		
Underweight versus normal	1.44 (0.84–2.47)	.182
Overweight versus normal	0.76 (0.67–0.87)	<.001
Obese versus normal	0.68 (0.58–0.79)	<.001
SF-36 Physical Component Score, standardised	0.77 (0.73–0.81)	<.001
SF-36 Mental Component Score, standardised	0.91 (0.87–0.96)	<.001
Number of medication codes	1.02 (1.00–1.05)	.026
Health behaviours: $\chi^2 = 46.80$, $p < .001$		
Cans of caffeinated soda/day	1.03 (0.97–1.09)	.365
Cups coffee or tea/day	1.02 (1.00–1.05)	.088
No alcohol use	1.17 (1.04–1.32)	.011
Smoking status: $\chi^2 = 33.87$, $p < .001$		
None versus current smoking	0.55 (0.45–0.67)	<.001
Past versus current smoking	0.66 (0.54–0.80)	<.001
Self-report sleep: $\chi^2 = 10.78$, $p = .029$		
Sleep duration: $\chi^2 = 1.56$, $p = .458$		
Short versus medium	0.93 (0.82–1.05)	.255
Long versus medium	1.02 (0.87–1.19)	.850
Frequency of difficulties getting back to sleep	0.96 (0.90–1.01)	.111
Frequency of excessive daytime sleepiness	0.95 (0.89–1.01)	.078
PSG sleep: $\chi^2 = 54.14$, $p < .001$		
Sleep efficiency: $\chi^2 = 12.42$, $p = .014$		
1st versus 2nd Quintile	1.05 (0.91–1.22)	.481
1st versus 3rd Quintile	1.26 (1.07–1.48)	.005
1st versus 4th Quintile	1.26 (1.06–1.50)	.010
1st versus 5th Quintile	1.06 (0.88–1.27)	.546
Stage 3%–4%: $\chi^2 = 4.96$, $p = .292$		
1st versus 2nd Quintile	1.14 (0.96–1.36)	.129
1st versus 3rd Quintile	0.96 (0.80–1.14)	.608
1st versus 4th Quintile	1.02 (0.85–1.23)	.831
1st versus 5th Quintile	0.97 (0.81–1.16)	.746
REM %: $\chi^2 = 18.39$, $p = .001$		
1st versus 2nd Quintile	1.27 (1.08–1.48)	.003

(Continues)

TABLE 2 (Continued)

Variable	HR (95% CI)	p
1st versus 3rd Quintile	1.32 (1.12–1.56)	<.001
1st versus 4th Quintile	1.18 (1.00–1.40)	.047
1st versus 5th Quintile	1.38 (1.16–1.65)	<.001
T90: $\chi^2 = 12.97$, $p = .002$		
Moderate versus low	1.13 (0.99–1.30)	.074
High versus low	1.28 (1.12–1.46)	<.001

BMI, body mass index; PSG, polysomnography; REM %, percentage of total sleep time in rapid eye movement sleep; SF-36, 36-item Short Form Health Survey; SHHS, Sleep Heart Health Study; Stage 3%–4%, percentage of total sleep time in Stages 3–4 non-REM sleep, corresponding with N3 or “slow-wave sleep”; T90, percentage of total sleep time spent with arterial oxygen saturation (SpO_2) <90%; WSCS, Wisconsin Sleep Cohort Study.

Importantly, the set of PSG-assessed measures was predictive of mortality across three different modelling strategies: machine learning with internal validation, machine learning with external validation, and Cox proportional hazards modelling.

Polysomnography measures of overnight hypoxaemia and SE ranked among the most predictive individual risk factors. Hypoxaemia and SE were significant in the random forest using external validation and the Cox proportional hazards model. However, they were not significant in the random forest using internal validation. The observation that the set of PSG measures was significant using internal validation, despite none of the individual measures it comprised being significant, highlights that simultaneous consideration of multiple PSG measures can enhance predictive accuracy.

The PRF approach used in the present study has much strength, including the ability to empirically model flexible relationships among variables and to provide unbiased rankings of predictive importance. However, because it does not provide traditional measures of effect size, supplementing the random forest with Cox proportional hazards modelling provides important additional information. Based on the Cox proportional hazards model, high T90 (>1.85%) and very low SE (<75%) were associated with the greatest risk for mortality. Hypoxaemia is most associated with sleep apnoea but may also identify sleep-related gas exchange abnormalities, including sleep-related hypoventilation and severe chronic obstructive pulmonary disease (Budhiraja, Siddiqi, & Quan, 2015; Dewan, Nieto, & Somers, 2015). Very low SE can occur with a wide range of sleep disturbances and disorders, including insomnia, SDB, and restless leg syndrome. It is also noteworthy that very low SE increased mortality risk relative to more moderate levels of SE, but not to very high SE. In some adults, very high SE may indicate sleep pathology in the form of increased sleep propensity, in association with increased daytime sleepiness (Wallace, Lee, et al., 2019), dampening the protective effects that may otherwise be expected.

Findings from our novel machine learning framework were generally consistent with those from the Cox model. However, only the Cox model indicated that REM % was a significant predictor of mortality. Prior publications using both regression and ad hoc random forest machine learning variable importance metrics (Dew et al., 2003; Leary et al., 2020; Zhang et al., 2019) have shown the importance of REM sleep for health and mortality. However, unlike the novel permutation variable importance metrics we present here, the ad hoc importance metrics used previously can be biased toward correlated predictors (Nicodemus et al., 2010; Strobl et al., 2007). In our novel machine learning approach, the predictive information within REM % is likely being explained by complex combinations of other risk factors, some of which (e.g. hypertension, diabetes) are potential or established intermediate factors in the causal pathway connecting REM % and mortality (Aurora, McGuffey, & Punjabi, 2020; Lecube et al., 2017; Mokhlesi et al., 2014). Similarly, we also observed that the health behaviours domain and several individual non-sleep predictors (e.g. BMI, smoking, alcohol use, number of prescription medications) were significant in the Cox model but not in the random forests.

Our present findings have important implications for risk screening. While it is unlikely that laboratory-based PSG would be used for wide-scale screening or health planning, the emergence of new, low-burden portable technologies for in-home sleep assessments make objective measurements of targeted dimensions of sleep feasible. The PSG measures that were most predictive (SE and T90) are relatively simple metrics that can be captured using single-channel electroencephalogram and overnight oximetry (e.g. see Areia et al., 2020; Lunsford-Avery et al., 2020). SE can also be captured using wearables that estimate sleep–wake time from accelerometry (e.g. see Kubala et al., 2020). Although the purpose of the present study was not to explicitly develop a risk-assessment tool incorporating PSG, our findings indicate such an algorithm is within reach and could be robust to differences in patient health profiles, measurement devices, and PSG scoring criteria. The algorithm could potentially be incorporated with built-in PSG software that automatically reports the scores and has potential to identify people at higher risk for additional screening or interventions. For example, individuals identified through such screening procedures as having the lowest SE and the highest T90 may be candidates for further assessment of insomnia, sleep apnoea, and cardio-pulmonary disorders.

Our present results also suggest that SE and T90 may be key targets in future novel mechanistic and treatment research. Findings from the animal literature suggest that hypoxaemia and low SE may elicit and maintain a physiological cascade that increases risk of adverse health outcomes (Dewan et al., 2015; Gaines, Vgontzas, Fernandez-Mendoza, & Bixler, 2018; Gileles-Hillel, Kheirandish-Gozal, & Gozal, 2016). Thus, new mechanistic studies and treatments might be most effectively directed to those with very low levels of SE, potentially targeting improvements in SE as a therapeutic goal for improving health outcomes. Individuals with high levels of overnight hypoxaemia may also be a priority of targeted interventions. However, in both cases, further individualised assessments to determine the specific aetiologies of reduced sleep efficiency (e.g. insomnia, SDB) or high hypoxaemia (SDB, pulmonary disease) would likely be needed. To this end, we also investigated the AHI, which more specifically reflects the intermittent nature of hypoxaemia and sleep fragmentation seen in SDB. However, for all-cause mortality, the less-specific T90 was preferable to the AHI as a predictor, perhaps because T90 can identify other pulmonary conditions and has more reliable scoring.

Self-reported sleep measures did notably not enhance prediction of mortality either on their own or in combination with PSG. However, other studies have indicated their potential importance (Gallicchio & Kalesan, 2009; Wallace, Buysse, et al., 2019; Wallace, Lee, et al., 2019). Thus, it will be important to differentiate profiles of individuals for whom objective sleep assessment, versus self-reported sleep assessment, is warranted, and to further evaluate alternative approaches for capturing predictive self-reported measurements.

Strengths of our present study include our application of a novel machine learning framework robust to complex inter-relationships among variables, the utilisation of two large well-characterised community-based epidemiological cohorts (the SHHS and WSCS), consideration of both physiological and self-reported measures of sleep, inclusion of a variety of established risk factors across several health domains, and up to 15 years of mortality follow-up. We also successfully used external validation (train in the SHHS, test in the WSCS) to show that PSG is predictive of mortality. This finding is notable given observed cohort differences between the SHHS and WSCS, including mortality rate and in-home versus laboratory-based PSG.

However, there are also study limitations to consider. First, the WSCS did not have enough deaths to be sufficiently powered for stand-alone analyses of the complexity performed here; such analyses might have indicated whether effects were moderated by key cohort differences. Second, because of the need to harmonise predictors across cohorts and restriction to measures captured within the existing cohorts, we could not include some established health/lifestyle behaviours (e.g. physical activity, diet, intensity of smoking, level of alcohol use). These measures could potentially have improved the predictive abilities of the health behaviours domain. Third, there were some differences in distributions of PSG measures between studies, which may be attributed to use of in-home versus laboratory PSG (Iber et al., 2004) and potential subtle differences in scoring procedures. Fourth, the findings are largely generalisable to White individuals in community settings, which limit our ability to extend findings to other important populations including racial/ethnic minorities.

In conclusion, physiological sleep measures, especially SE and T90, provide unique predictive information for all-cause mortality at levels on a par with risk factors routinely used for screening in primary care. These findings, coupled with the emergence of new low-burden technologies for objective sleep assessments, have important implications for risk screening and guiding treatment and mechanistic studies. In future research, it will be important to examine differences by age, sex, and cause-specific mortality outcomes; differentiate profiles of individuals for whom PSG versus self-report sleep assessment is warranted; and utilise more racially and ethnically diverse samples.

ACKNOWLEDGMENTS

The collection of the Wisconsin Sleep Cohort Study (WSCS) data is supported by the National Heart, Lung, and Blood Institute (NHLBI) [R01HL62252], National Institute of Aging [1R01AG036838, R01AG14124], and the National Center for Research Resources [1UL1RR025011] at the United States National Institutes of Health (NIH). The Sleep Heart Health Study (SHHS) is supported by the NHLBI [U01HL53916, U01HL53931, U01HL53934, U01HL53937, U01HL53938, U01HL53940, U01HL53941, U01HL64360]. This work was also supported by the NIH [R01AG056331 to MW, R35HL135818 to SR, R01AG047139 to Buysse/Hall] and the National Sleep Research Resource (NSRR) funded by the NHLBI [HL114473].

CONFLICT OF INTERESTS

DJB has served as a paid consultant to Bayer, BeHealth Solutions, Emmi Solutions, Pear Therapeutics, Sleep Number Corporation, and Weight Watchers International (past 5 years). He has served as a paid consultant for professional educational programmes developed by the American Academy of Physician Assistants, CME Institute and Emmi Solutions, and received payment for a professional education programme sponsored by Eisai. He is an author of the Pittsburgh Sleep Quality Index, Daytime Insomnia Symptoms Scale, Pittsburgh Sleep Diary, Insomnia Symptoms Questionnaire (copyright held by University of Pittsburgh). These instruments have been licensed to commercial entities for fees. He is also co-author of the Consensus Sleep Diary (copyright held by Ryerson University), which is licensed to commercial entities for a fee. He has received grant support from NIH, Patient-Centered Outcomes Research Institute (PCORI), Agency for Healthcare Research and Quality (AHRQ), and the Veteran's Administration (VA). MLW served as a paid consultant for Noctem and has received grant support from the NIH. KLS reports grants from NIH and Merck & Co. SR has received grant support from the NIH, reports a grant from Jazz Pharma and consulting fees from Jazz Pharma and Respicardia. MHH, JLG, LKM, TSC, PEP, and EWH have no conflict of interest to declare.

AUTHOR CONTRIBUTIONS

MLW conceptualised the study, analysed data, and drafted the manuscript. TSC, LKM, and JLG assisted with data analysis and editing. DJB, EWH, MHH, KLS, RS, and PEP assisted in study conceptualisation, interpretation, and editing.

DATA AVAILABILITY STATEMENT

Some SHHS and WSCS data analysed in the present study are available through the National Sleep Research Resource. Statistical code is available on request.

ORCID

Meredith L. Wallace https://orcid.org/0000-0003-3951-890X

REFERENCES

Areia, C., Young, L., Vollam, S., Ede, J., Santos, M., Tarassenko, L., & Watkinson, P. (2020). Wearability testing of ambulatory vital sign monitoring devices: prospective observational cohort study. *JMIR Mhealth Uhealth*, 8(12), e20214. https://doi.org/10.2196/20214

Aurora, R. N., McGuffey, E. J., & Punjabi, N. M. (2020). Natural history of sleep-disordered breathing during rapid eye movement sleep. relevance for incident cardiovascular disease. *Annals of the American Thoracic Society*, *17*(5), 614–620. https://doi.org/10.1513/AnnalsATS.201907-524OC

Budhiraja, R., Siddiqi, T. A., & Quan, S. F. (2015). Sleep disorders in chronic obstructive pulmonary disease: Etiology, impact, and management. *Journal of Clinical Sleep Medicine*, *11*(3), 259–270. https://doi.org/10.5664/jcsm.4540

Buysse, D. J. (2014). Sleep health: can we define it? Does it matter? *Sleep*, *37*(1), 9–17. https://doi.org/10.5665/sleep.3298

Cao, J., & Zhang, S. (2014). Multiple comparison procedures. *JAMA*, *312*(5), 543–544. https://doi.org/10.1001/jama.2014.9440

Coleman, T., Peng, W., & Mentch, L. (2019). Scalable and efficient hypothesis testing with random forests. *arXiv*, arXiv:1904.07830.

Collins, G. S., Ogundimu, E. O., & Altman, D. G. (2016). Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statistics in Medicine*, *35*(2), 214–226. https://doi.org/10.1002/sim.6787

Dean, D. A., Goldberger, A. L., Mueller, R., Kim, M., Rueschman, M., Mobley, D., ... Redline, S. (2016). Scaling up scientific discovery in sleep medicine: The National Sleep Research Resource. *Sleep*, *39*(5), 1151–1164. https://doi.org/10.5665/sleep.5774.

Dew, M. A., Hoch, C. C., Buysse, D. J., Monk, T. H., Begley, A. E., Houck, P. R., ... Reynolds, C. F. (2003). Healthy older adults' sleep predicts all-cause mortality at 4 to 19 years of follow-up. *Psychosomatic Medicine*, *65*(1), 63–73.

Dewan, N. A., Nieto, F. J., & Somers, V. K. (2015). Intermittent hypoxemia and OSA: implications for comorbidities. *Chest*, *147*(1), 266–274. https://doi.org/10.1378/chest.14-0500

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, *15*(1), 3133–3181.

Gaines, J., Vgontzas, A. N., Fernandez-Mendoza, J., & Bixler, E. O. (2018). Obstructive sleep apnea and the metabolic syndrome: The road to clinically-meaningful phenotyping, improved prognosis, and personalized treatment. *Sleep Medicine Reviews*, *42*, 211–219. https://doi.org/10.1016/j.smrv.2018.08.009

Gallicchio, L., & Kalesan, B. (2009). Sleep duration and mortality: a systematic review and meta-analysis. *Journal of Sleep Research*, *18*(2), 148–158. https://doi.org/10.1111/j.1365-2869.2008.00732.x

Gileles-Hillel, A., Kheirandish-Gozal, L., & Gozal, D. (2016). Biological plausibility linking sleep apnoea and metabolic dysfunction. *Nat Rev Endocrinol*, *12*(5), 290–298. https://doi.org/10.1038/nrendo.2016.22

Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics*, *61*(1), 92–105. https://doi.org/10.1111/j.0006-341X.2005.030814.x

Hori, T., Sugita, Y., Koga, E., Shirakawa, S., Inoue, K., Uchida, S., ... Fukuda, N. (2001). Proposed supplements and amendments to 'A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects', the Rechtschaffen & Kales (1968) standard. *Psychiatry and Clinical Neurosciences*, *55*(3), 305–310. https://doi.org/10.1046/j.1440-1819.2001.00810.x

Iber, C., Redline, S., Gilpin, A. M. K., Quan, S. F., Zhang, L., Gottlieb, D. J., ... Smith, P. (2004). Polysomnography performed in the unattended home versus the attended laboratory setting–Sleep Heart Health Study methodology. *Sleep*, *27*(3), 536–540. https://doi.org/10.1093/sleep/27.3.536

Kubala, A. G., Barone Gibbs, B., Buysse, D. J., Patel, S. R., Hall, M. H., & Kline, C. E. (2020). Field-based measurement of sleep: Agreement between six commercial activity monitors and a validated accelerometer. *Behavioral Sleep Medicine*, *18*(5), 637–652.

Leary, E. B., Watson, K. T., Ancoli-Israel, S., Redline, S., Yaffe, K., Ravelo, L. A., ... Stone, K. L. (2020). Association of rapid eye movement sleep with mortality in middle-aged and older adults. *JAMA Neurology*, *77*(10), 1–12. https://doi.org/10.1001/jamaneurol.2020.2108

Lecube, A., Romero, O., Sampol, G., Mestre, O., Ciudin, A., Sánchez, E., ... Simó, R. (2017). Sleep biosignature of Type 2 diabetes: a case-control study. *Diabetic Medicine*, *34*(1), 79–85. https://doi.org/10.1111/dme.13161

Lunsford-Avery, J. R., Keller, C., Kollins, S. H., Krystal, A. D., Jackson, L., & Engelhard, M. M. (2020). Feasibility and acceptability of wearable sleep electroencephalogram device use in adolescents: Observational Study. *JMIR Mhealth Uhealth*, *8*(10), e20590.

Mentch, L., & Hooker, G. (2017). Formal hypothesis tests for additive structure in random forests. *Journal of Computational and Graphical Statistics*, *26*(3), 589–597.

Mokhlesi, B., Finn, L. A., Hagen, E. W., Young, T., Hla, K. M., Van Cauter, E., & Peppard, P. E. (2014). Obstructive sleep apnea during REM sleep and hypertension. results of the Wisconsin Sleep Cohort. *American Journal of Respiratory and Critical Care Medicine*, *190*(10), 1158–1167. https://doi.org/10.1164/rccm.201406-1136OC

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*, 110. https://doi.org/10.1186/1471-2105-11-110

Peppard, P. E., Young, T., Barnet, J. H., Palta, M., Hagen, E. W., & Hla, K. M. (2013). Increased prevalence of sleep-disordered breathing in adults. *American Journal of Epidemiology*, *177*(9), 1006–1014. https://doi.org/10.1093/aje/kws342

Punjabi, N. M., Caffo, B. S., Goodwin, J. L., Gottlieb, D. J., Newman, A. B., O'Connor, G. T., ... Samet, J. M. (2009). Sleep-disordered breathing and mortality: A prospective cohort study. *PLoS Med*, *6*(8), e1000132. https://doi.org/10.1371/journal.pmed.1000132

Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., ... Wahl, P. W. (1997). The Sleep Heart Health Study: design, rationale, and methods. *Sleep*, *20*(12), 1077–1085.

Rechtschaffen, A., & Kales, A. (1968). *A manual of standardized techniques and scoring system for sleep stages of human subjects*. Washington, DC: United States Government Printing Office

Redline, S., Kirchner, H. L., Quan, S. F., Gottlieb, D. J., Kapur, V., & Newman, A. (2004). The effects of age, sex, ethnicity, and sleep-disordered breathing on sleep architecture. *Archives of Internal Medicine*, *164*(4), 406–418. https://doi.org/10.1001/archinte.164.4.406

Redline, S., Sanders, M. H., Lind, B. K., Quan, S. F., Iber, C., Gottlieb, D. J., ... Kiley, J. P. (1998). Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. *Sleep Heart Health Research Group. Sleep*, *21*(7), 759–767.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior*, *29*(5), 615–620. https://doi.org/10.1007/s10979-005-6832-7

Senaratna, C. V., Perret, J. L., Lodge, C. J., Lowe, A. J., Campbell, B. E., Matheson, M. C., ... Dharmage, S.C. (2017). Prevalence of obstructive sleep apnea in the general population: A systematic review. *Sleep Medicine Reviews*, *34*, 70–81. https://doi.org/10.1016/j.smrv.2016.07.002

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, *8*, 25. https://doi.org/10.1186/1471-2105-8-25

Swenson, P. F., & Ebell, M. H. (2016). Introducing a one-page adult preventive health care schedule: USPSTF recommendations at a glance. *American Family Physician*, *93*(9), 738–740.

Wallace, M. L., Buysse, D. J., Redline, S., Stone, K. L., Ensrud, K., Leng, Y., … Hall, M. H. (2019). Multidimensional sleep and mortality in older adults: A machine-learning comparison with other risk factors. *Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, *74*(12), 1903–1909. https://doi.org/10.1093/gerona/glz044

Wallace, M. L., Lee, S., Hall, M. H., Stone, K., Langsetmo, L., & Redline, S. … MrOS and SOF Research Groups. (2019). Heightened sleep propensity: A novel and high-risk sleep health phenotype in older adults. *Sleep Health*, *5*(6), 630–638. https://doi.org/10.1016/j.sleh.2019.08.001

Wallace, M. L., Stone, K., Smagula, S. F., Hall, M. H., Simsek, B., Kado, D. M., … Buysse, D. J. (2018). Which sleep health characteristics predict all-cause mortality in older men? An application of flexible multivariable approaches. *Sleep*, *41*(1), zsx189. https://doi.org/10.1093/sleep/zsx189

Ware, J., Kosinski, M., & Gandek, B. (2000). *SF-36 Health Survey: Manual and Interpretation Guide*. Lincoln, RI: QualityMetric Inc.

Whitney, C.W., Gottlieb, D. J., Redline, S., Norman, R. G., Dodge, R. R., Shahar, E., … Nieto, F.J. (1998). Reliability of scoring respiratory disturbance indices and sleep staging. *Sleep*, *21*(7), 749–757. https://doi.org/10.1093/sleep/21.7.749

Young, T., Finn, L., Peppard, P. E., Szklo-Coxe, M., Austin, D., Nieto, F. J., … Hla, K. M. (2008). Sleep disordered breathing and mortality: eighteen-year follow-up of the Wisconsin sleep cohort. *Sleep*, *31*(8), 1071–1078.

Young, T., Palta, M., Dempsey, J., Skatrud, J., Weber, S., & Badr, S. (1993). The occurrence of sleep-disordered breathing among middle-aged adults. *New England Journal of Medicine*, *328*(17), 1230–1235. https://doi.org/10.1056/nejm199304293281704

Zhang, G. -Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., … Redline, S. (2018). The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, *25*(10), 1351–1358. https://doi.org/10.1093/jamia/ocy064

Zhang, J., Jin, X., Li, R., Gao, Y., Li, J., & Wang, G. (2019). Influence of rapid eye movement sleep on all-cause mortality: a community-based cohort study. *Aging*, *11*(5), 1580–1588. https://doi.org/10.18632/aging.101858

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Wallace ML, Coleman TS, Mentch LK, et al. Physiological sleep measures predict time to 15-year mortality in community adults: Application of a novel machine learning framework. *J Sleep Res*. 2021;30:e13386. https://doi.org/10.1111/jsr.13386