

## ARTICLE OPEN



# Age estimation from sleep studies using deep learning predicts life expectancy

Andreas Brink-Kjaer<sup>1,2,3</sup>, Eileen B. Leary<sup>3</sup>, Haoqi Sun<sup>4</sup>, M. Brandon Westover<sup>4</sup>, Katie L. Stone<sup>5,6</sup>, Paul E. Peppard<sup>7</sup>, Nancy E. Lane<sup>8</sup>, Peggy M. Cawthon<sup>5,6</sup>, Susan Redline<sup>9,10</sup>, Poul Jennum<sup>2,11</sup>, Helge B. D. Sorensen<sup>1,11</sup> and Emmanuel Mignot<sup>3,11</sup>

Sleep disturbances increase with age and are predictors of mortality. Here, we present deep neural networks that estimate age and mortality risk through polysomnograms (PSGs). Aging was modeled using 2500 PSGs and tested in 10,699 PSGs from men and women in seven different cohorts aged between 20 and 90. Ages were estimated with a mean absolute error of  $5.8 \pm 1.6$  years, while basic sleep scoring measures had an error of  $14.9 \pm 6.29$  years. After controlling for demographics, sleep, and health covariates, each 10-year increment in age estimate error (AEE) was associated with increased all-cause mortality rate of 29% (95% confidence interval: 20–39%). An increase from  $-10$  to  $+10$  years in AEE translates to an estimated decreased life expectancy of 8.7 years (95% confidence interval: 6.1–11.4 years). Greater AEE was mostly reflected in increased sleep fragmentation, suggesting this is an important biomarker of future health independent of sleep apnea.

npj Digital Medicine (2022)5:103; <https://doi.org/10.1038/s41746-022-00630-9>

## INTRODUCTION

Sleep clinics throughout the world evaluate millions of patients every year. The gold standard diagnostic test for this evaluation is nocturnal polysomnography (PSG), a test comprised of multiple physiological signals, i.e., electroencephalogram (EEG), electrocardiogram (ECG), electrooculogram (EOG), chin and leg electromyogram (EMG), breathing effort and airflow, all of which are recorded overnight. The PSG provides recording of multiple physiological measures during sleep, at a time when the individual is mostly immobile and uncontaminated by sensory inputs. It thus contains a wealth of information on the normal physiology of a given individual (notably brain physiology).

Sadly, the millions of PSGs collected every year are primarily used clinically to visually extract simple metrics such as sleep latency, proportion of time in various sleep stages, rates of sleep apnea events (apnea-hypopnea index, AHI), periodic leg movement (PLM), and arousals (arousal index, Ari). Scoring is done manually by trained technicians and supervised by medical doctors, according to American Academy of Sleep Medicine (AASM) guidelines<sup>1</sup>. This scoring is time-consuming and prone to inter- and intra-rater variability<sup>2</sup>. Of particular clinical importance are measures of sleep disordered breathing events such as the AHI or associated hypoxic burden, which has been associated with daytime sleepiness<sup>3</sup>, cognitive impairment, and increased risk of cardiovascular disease such as development of high blood pressure and stroke in multiple studies independent of age, sex and obesity<sup>4–9</sup>. Sleep apnea has also been shown to be associated with increased mortality risk independent of obesity, age, and sex<sup>10</sup>.

Although sleep apnea measures are currently the main rationale for conducting clinical sleep studies, there is evidence that other aspects of objective sleep influence mortality and health outcomes. All-cause mortality has been associated with an increase in

arousal burden<sup>11</sup> (a measure of sleep fragmentation), decreased sleep efficiency (SE)<sup>12</sup> and decreased rapid eye movement (REM) sleep amounts<sup>13</sup>. Similarly, decreased slow-wave sleep and low SE have been associated with hypertension incidence and a variety of cardiovascular outcomes among participants in the Sleep Heart Health Study (SHHS)<sup>14,15</sup>. Finally, specific abnormalities such as REM sleep behavior disorder (RBD) and loss of sleep-stage specific autonomic regulation during sleep are well established early precursors of synucleinopathies<sup>16–18</sup>.

Recently, promising deep learning methods have been developed that efficiently and objectively assist PSG analyses<sup>19–21</sup>. These algorithms provide added information such as higher resolution sleep stages and probabilistic measures, in contrast to manual scoring that only offers categorical classification. However, these new methods have mostly been confined to replicating a scoring practice that is limited by arbitrary definitions<sup>1</sup> that may not capture all relevant information available in the data. Further, they merely imitate human scoring without attempting to capture all the rich incipient information contained in a full night PSG study discussed above. Deep learning methods that utilize all relevant information in PSGs may provide additional useful clinical insights such as important health outcomes.

Age is one of the strongest predictors of morbidity and mortality. Sleep architecture and subjective sleep complaints are also affected by aging<sup>22,23</sup>. As people age, sleep becomes shorter<sup>23</sup>, more fragmented<sup>24</sup>, exhibits fewer sleep spindles<sup>25</sup>, includes less slow wave sleep, and, to a lesser extent, less REM sleep<sup>26</sup>. Moreover, several of these changes have been linked to increased mortality, even after controlling for the effects of age<sup>11–13</sup>.

A recent study modeled the age of subjects based on automatic sleep-stage features from EEG recordings<sup>27</sup>. Furthermore, this model's age estimate (AE) error (AEE), the model residual that represented a brain aging index, was associated with increased

<sup>1</sup>Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark. <sup>2</sup>Danish Center for Sleep Medicine, Department of Clinical Neurophysiology, Rigshospitalet, Denmark. <sup>3</sup>Stanford Center for Sleep Sciences and Medicine, Stanford University, Palo Alto, CA, USA. <sup>4</sup>Department of Neurology, Massachusetts General Hospital, Boston, MA, USA. <sup>5</sup>Research Institute, California Pacific Medical Center, San Francisco, CA, USA. <sup>6</sup>Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA. <sup>7</sup>Department of Population Health Sciences, University of Wisconsin-Madison, Madison, WI, USA. <sup>8</sup>Department of Medicine, University of Davis School of Medicine, Sacramento, CA, USA. <sup>9</sup>Department of Medicine, Harvard Medical School, Boston, MA, USA. <sup>10</sup>Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. <sup>11</sup>These authors contributed equally: Poul Jennum, Helge B. D. Sorensen, Emmanuel Mignot. ✉email: [andbri@dtu.dk](mailto:andbri@dtu.dk); [mignot@stanford.edu](mailto:mignot@stanford.edu)

**Table 1.** Mean absolute error of age estimation models.

Model	MAE			
	Train set $n = 2500$	Val set $n = 200$	Test set $n = 10,509$	HomePAP* $n = 190$
Basic sleep measures	14.9 ± 6.08	14.9 ± 6.53	14.6 ± 5.91	12.5 ± 4.06
(a) Central EEG	5.43 ± 1.25	6.52 ± 2.48	6.77 ± 2.2	7.65 ± 2.7
(b) EEG+EOG+EMG	5.35 ± 0.96	5.88 ± 2.09	6.81 ± 1.84	8.62 ± 2.92
(c) ECG	9.11 ± 1.89	11 ± 4.05	10.4 ± 2.23	13.9 ± 6.74
(d) Respiratory	8.87 ± 2.2	9.31 ± 2.39	8.09 ± 1.89	13.7 ± 6.05
(e) Ensemble–Avg.	5.4 ± 1.01	6.11 ± 1.84	5.8 ± 1.16	8.16 ± 3.75

The MAE is reported as mean ± standard deviation and was averaged across age intervals ([20, 25], [25, 30], ..., [85–89]), which are reported for the test and HomePAP set in Supplementary Tables 2 and 3. \*The training and validation set includes no PSGs from the HomePAP study, thus it represents expected performance in a new unseen cohort with a different technical setup. Basic sleep measures denote a linear regression model with the following predictive variables: arousal index, apnea-hypopnea index, total sleep time, wake after sleep onset, and percentage of N1, N2, N3, and REM sleep. MAE: mean absolute error.

risk of mortality<sup>28</sup>, dementia<sup>29</sup>, and human immunodeficiency virus (HIV)<sup>30</sup>. However, as the authors pointed out, this approach was still limited by the use of hand-crafted features, and used only the EEG signal, whereas other physiological signals also carry important information about health and life expectancy. Nonetheless, since age is readily available in all subjects, (unlike mortality or other outcomes) predicting age may be a reasonable first proxy to predicting poor outcomes in a variety of disease area.

In this study, we built on this previous study aiming by (1) modeling age, as a proxy for mortality risk, directly using deep learning models; (2) interpreting the features learned by the models; and (3) investigating associations between the AEE of the models and both all-cause and cardiovascular mortality.

## RESULTS

### Performance of age estimation models

In this study, we used a combined sample of 13,332 PSGs from seven cohorts: the Stanford Technology Analytics and Genomics of Sleep (STAGES)<sup>31,32</sup>, the Stanford Sleep Cohort (SSC)<sup>33,34</sup>, the Wisconsin Sleep Cohort (WSC)<sup>4,34</sup>, the SHHS<sup>35,36</sup>, the Osteoporotic Fractures in Men (MrOS) Sleep Study<sup>36–38</sup>, the Cleveland Family Study (CFS)<sup>36,39</sup>, and the Home Positive Airway Pressure (HomePAP) Study<sup>36,40</sup>.

A set of AE models, comprised of deep neural networks, were trained on 2500 PSGs from subjects with a close to uniform age distribution between 6 and 90 years. These AE models each used a set of input PSG signals: (a, Central EEG) C3-M2, C4-M1; (b, EEG + EOG + EMG) C3-M2, C4-M1, L-EOG, R-EOG, chin EMG; (c, ECG) ECG; (d, respiratory) airflow, nasal pressure, thoracic and abdominal belts, blood oxygen saturation. Finally, an ensemble model (e, Ensemble–Avg.) was developed based on the average AE of models (a), (b), (c), and (d). A validation set of 200 PSGs were used to optimize hyperparameters of the AE models, of which the final hyperparameter tunings are shown in Supplementary Table 1.

Performance of various AE models (based on EEG alone or various components of the PSG, see Table 1) was evaluated as mean absolute error (MAE) stratified by 5-year age intervals, as shown in Supplementary Table 2 for the first test set, and in Supplementary Table 3 for the HomePAP study (a second test dataset with an age range from 20 to 80 years). The stratification weighs each age interval equally despite a non-uniform age distribution. Table 1 shows MAE for each data subset averaged across all 5-year age groups ranging from 20 to 90 years. The best performing model on the test set was the (e, Ensemble–Avg.) model, which averages model (a–d), while the (a, Central EEG) model generalized best to the HomePAP dataset. As a comparison, we also report performance of using basic sleep study

metrics for age estimation, which includes Ari, AHI, total sleep time (TST), wake after sleep onset (WASO), and percentage of NREM stages (N1, N2, N3), and REM sleep.

A scatterplot of AE for model (e, Ensemble–Avg.) and chronological age for the test set and HomePAP data is shown in Fig. 1.

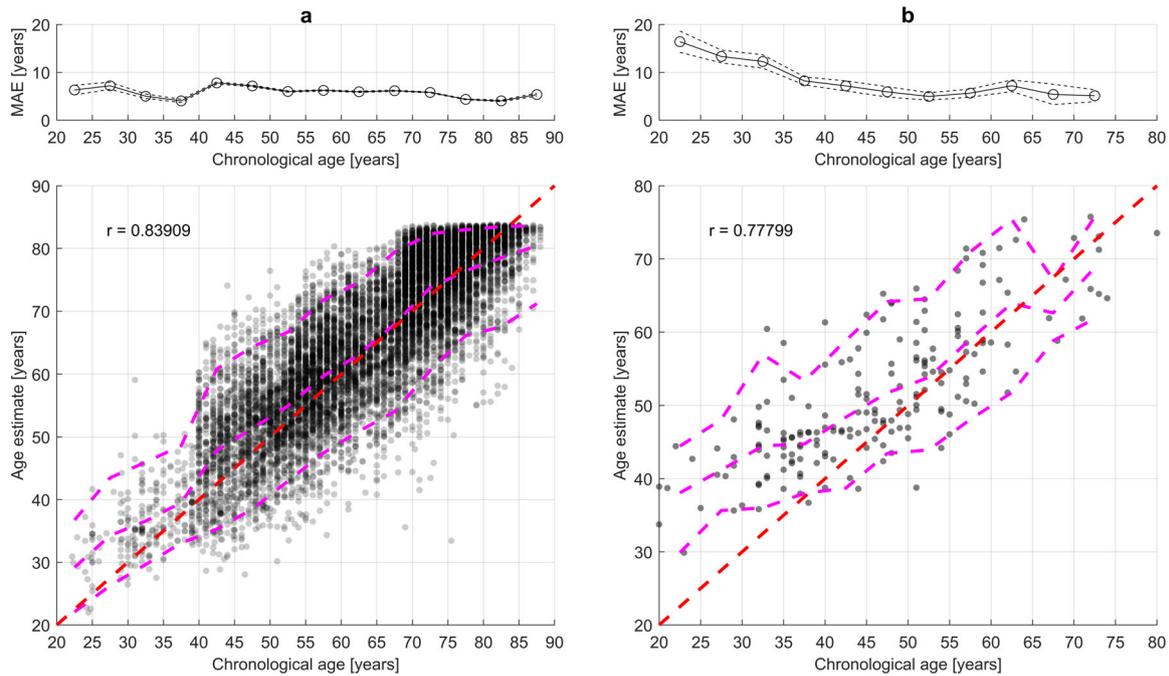
Night-to-night variability was investigated in the STAGES dataset ( $n = 42$ ). MAE was 5.93 years and 7.31 years during night 1 and 2, respectively. The difference between night 2 and 1 was  $-1.17 \pm 5.71$  (mean ± standard deviation), which was not significantly different from 0 ( $p = 0.19$ ). The absolute difference between nights were  $4.42 \pm 3.74$  years ( $p = 2.10^{-9}$ ).

The reliability of the AEs in longitudinal data was investigated in the WSC ( $n = 505$ ) with a time of  $4.08 \pm 1.02$  years between visits. The MAE was  $4.34 \pm 3.07$  years for the first visit and  $4.51 \pm 3.32$  years for the second visit. The AE increased by  $3.37 \pm 4.05$  between visits. Hence, the average increase was 0.7 years higher than for the chronological age ( $p = 0.00016$ ).

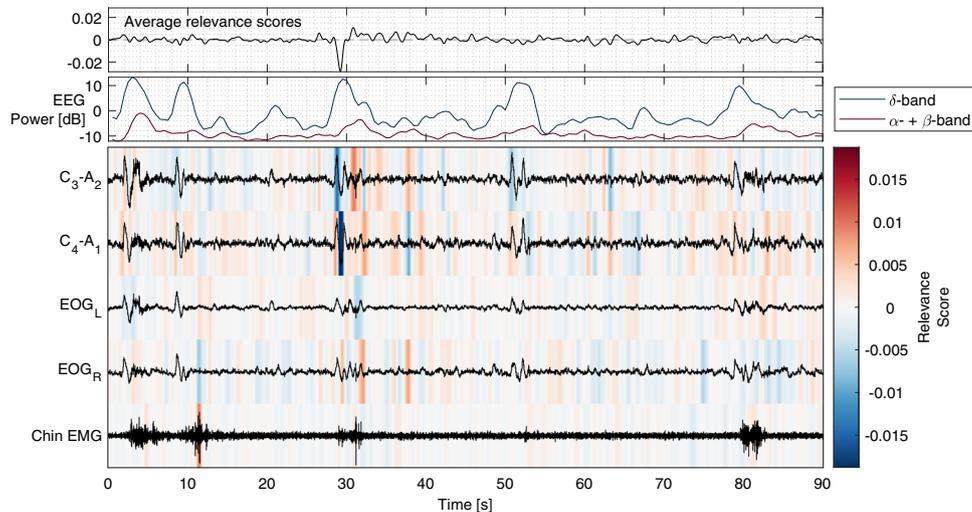
### Interpretation of deep learning framework and of age estimation errors

The age difference obtained between the various AE models and chronological age, i.e., AEE, can be considered a measure of how much “younger” or “older” sleep in a PSG appears. As a sanity check, we first examined associations between AEE of the models with basic sleep measures, which are shown in Supplementary Table 4. In general, higher AEE was associated with worse sleep based on metrics related to sleep fragmentation [Ari, SE, WASO, TST, and N1%]. The respiratory-based AEE shows a very strong association with the AHI ( $b = 1.5$ ,  $p = 4.7 \cdot 10^{-76}$ ), suggesting that it indeed captures information about sleep disordered breathing, which is known to increase with age.

Associations between AEE and sex, body mass index (BMI), medication use (antidepressants and benzodiazepines), and morbidities [hypertension, history of heart attack, congestive heart failure (CHF), chronic obstructive pulmonary disease (COPD), type 2 diabetes (T2D), and stroke] are shown in Supplementary Table 5. Presence of T2D was associated with a higher AEE ( $b = 1.6$ ,  $p = 9.0 \cdot 10^{-7}$ ) for the (a, Central EEG) model and ( $b = 1.2$ ,  $p = 7.7 \cdot 10^{-5}$ ) for the (b, EEG + EOG + EMG) model. For the (c, ECG) model, all heart related comorbidities were associated with a higher AEE (hypertension: 2.2 years,  $p = 8.8 \cdot 10^{-24}$ ; CHF: 3.1 years,  $p = 5.1 \cdot 10^{-7}$ ; history of heart attack: 1.8 years,  $p = 3.3 \cdot 10^{-6}$ ). Moreover, hypertension was associated with higher AEE in all but the (d, respiratory) model. Sex and BMI was associated with higher AEE in the (d, respiratory) model (sex:  $b = 3.6$ ,  $p = 3.5 \cdot 10^{-96}$ ; BMI:  $b = 1.2$ ,  $p = 2.0 \cdot 10^{-54}$ ). As for the stroke, COPD, and use of benzodiazepines, no significant associations to AEE were found.



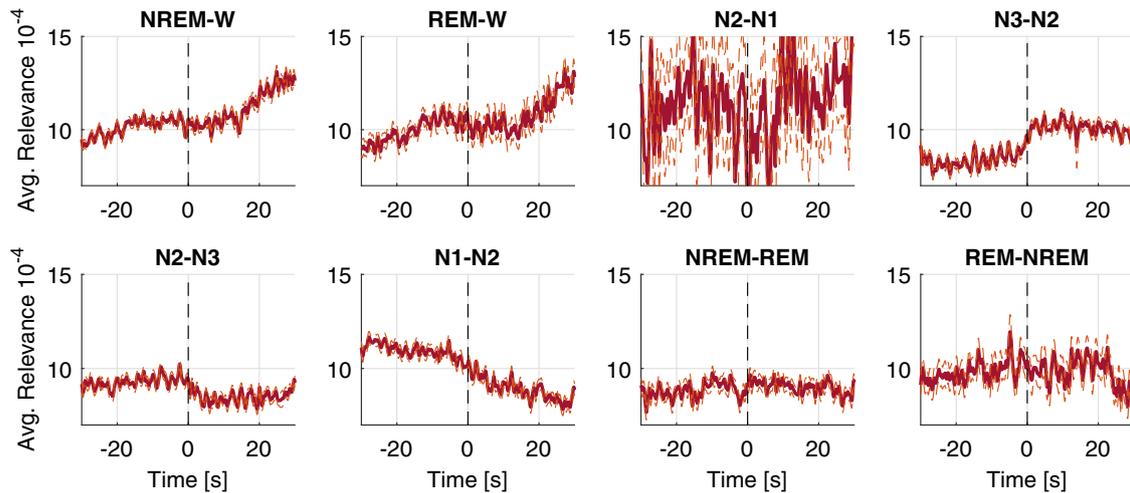
**Fig. 1** Scatterplot of age estimate and chronological age in the test sets for model (e, Ensemble-Avg.). **a** The test set ( $n = 9899$ ). The dotted line indicates the standard error of the mean (SEM) calculated as  $\sigma/\sqrt{n}$ . **b** The HomePAP test set ( $n = 190$ ). The red line indicates the optimal age estimate; the magenta lines indicate 5th, 50th, and 95th percentiles of age estimate in 5-year intervals.  $r$  is Pearson's correlation coefficient between age estimate and chronological age. MAE: mean absolute error.



**Fig. 2** Example of model (b; EEG+EOG+EMG) interpretation through relevance attribution of samples. The top plot shows relevance scores averaged across channels (C3-A2, C3-A1, EOG<sub>L</sub>, EOG<sub>R</sub>, Chin EMG). The second plot shows EEG power in the  $\delta$ -band (0–4 Hz) and the combined  $\alpha$ - and  $\beta$ -bands (>8 Hz). Red and blue indicate positive and negative attribution to the age estimate, respectively. Relevance attribution was computed using gradient SHAP.

Gradient SHAP<sup>41,42</sup> (SHapley Additive exPlanations) was used to attribute relevance scores of the AE to input PSG signal samples. For a given PSG, each signal sample was attributed with a relevance score that add up the AE for all samples in that PSG. Visual interpretation of relevance attribution, as shown in Fig. 2, shows that model (b, EEG + EOG + EMG) AE is increased in the presence of arousals and decreased in the presence of slow-wave oscillations. Furthermore, as shown in Supplementary Fig. 1, model (d, respiratory) AE is elevated in the presence of sleep apnea, and model (c, ECG) AE indicates that arrhythmias contribute to its AE.

The relationship between relevance scores and manually scored sleep events was investigated to validate that these are meaningful to the AE models. Relevance scores were averaged around transitions of manually scored hypnograms, arousal, and apnea/hypopnea events in PSGs from the CFS, the MrOS, and SHHS cohort in the training set. In Fig. 3, relevance scores of model (b; EEG + EOG + EMG) time-locked to sleep-stage transitions are shown. On average, the relevance scores of model (b; EEG + EOG + EMG) are increased when transitioning to lighter sleep or wakefulness. Furthermore, as shown in Supplementary Fig. 2, the average relevance scores are affected by arousal and apnea.



**Fig. 3 Average and smoothed relevance attribution averaged over channels of model (b; EEG+EOG+EMG) time-locked to sleep-stage transitions.** The average relevance attribution time-locked to sleep-stage transitions. These were averaged in 4353 PSGs from the test set with available manual scoring. The dotted line marks the standard error of the mean. Relevance attribution was computed using gradient SHAP.

### Association between age estimate error and mortality

Older age is the major predictor of mortality, an obvious application of our AEE calculation was to explore whether a positive AEE predicts increased mortality.

The combined dataset of subjects with both a PSG and associated mortality data consisted of 9386 subjects from the SHHS ( $n = 5696$ , deaths = 1285), the MrOS ( $n = 2781$ , deaths = 1662), and the WSC ( $n = 909$ , deaths = 98). This subset of data was also used in the training, validation, and test set for age estimation. The combined sample of subjects had a mean age of  $66.0 \pm 11.1$  years at baseline and was followed for a median of  $12.1 \pm 3.7$  years.

Supplementary Table 6 shows the association between all-cause mortality and a set of demographic, lifestyle, and health characteristics that we investigated with Cox proportional hazard models adjusted for age, sex, BMI, and cohort. The table also displays the proportion of missing data, which was imputed for further analyses.

The distributions of all demographics, lifestyle, and health characteristics across quartiles of the corrected AEE (AEEc, which is AEE corrected for age bias) for model (e, Ensemble-Avg.) are shown in Supplementary Tables 7–9 for the SHHS, WSC, and MrOS, respectively. Most notably, hypertension was more prevalent in the highest AEEc quartile.

After controlling for covariates (see Table 2), each 10-year increment in the AEE of model (e, Ensemble-Avg.), of which the standard deviation was 6.82 years in this combined dataset, was associated with a 29% (HR = 1.29, 95% confidence interval [CI]: 1.20–1.39) and 40% (HR = 1.40, 95% CI: 1.21–1.62) increase in all-cause and cardiovascular mortality rates, respectively. In Supplementary Tables 10–12, the results of the mortality analyses in each cohort are shown. Restricting the analyses to individual cohorts revealed that the association between AEE and mortality is present in the SHHS and MrOS cohort, while analysis in the WSC yielded mostly non-significant effects. However, this could be explained by a lower sample size and fewer deaths in the WSC ( $n = 909$ , deaths = 98) compared to the other cohorts.

In Fig. 4, survival curves for an AEE of +10 and –10 years for model (e, Ensemble-Avg.) is shown, which was generated using Cox Model 3 with all other covariates are set to their mean value. The survival curve was extended to compute the change in life expectancy for a change in AEE from –10 to 10 years. For model (e, Ensemble-Avg.), given an age of 40, 60, or 80 years in Cox

Model 3, a decrease in life expectancy was 12.6 years (CI: 8.9–16.2), 8.7 years (CI: 6.1–11.4), or 6.0 years (CI: 4.2–7.8), respectively.

Because hypertension and sleep apnea were very common in these cohorts, we also examined the mortality association in subjects without hypertension and without sleep apnea (AHI  $\geq 15$ ). A sensitivity analysis (see Supplementary Table 13) found that isolating the analyses to a subset of subjects without hypertension ( $n = 5303$ , deaths = 1291) decreased the hazard ratios of increased AEE to (HR = 1.25, 95% CI: 1.11–1.40) and (HR = 1.31, 95% CI: 1.03–1.66) for all-cause and cardiovascular mortality, respectively. As shown in Supplementary Table 14, isolating the analyses to a subset of subjects without sleep apnea ( $n = 5161$ , deaths = 1390) decreased the hazard ratios of increased AEE to (HR = 1.22, 95% CI: 1.10–1.37) and (HR = 1.24, 95% CI: 1.01–1.54) for all-cause and cardiovascular mortality, respectively. These effects are significant within the 95% CI in both the hypertension and sleep apnea sensitivity analyses. Lastly, to justify the inclusion of training and validation data for our AE models, we restricted the analysis to the test set ( $n = 8432$ , deaths = 2601). As shown in Supplementary Table 15, the hazard ratios of AEE are slightly decreased to (HR = 1.27, 95% CI: 1.18–1.38) and (HR = 1.35, 95% CI: 1.16–1.56) for all-cause and cardiovascular mortality, respectively.

### DISCUSSION

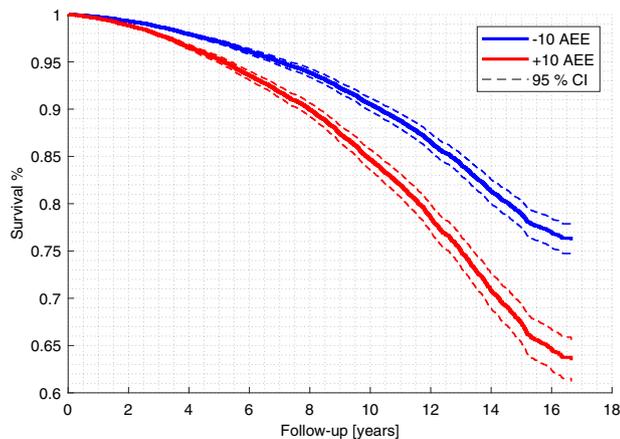
Our results show that deep learning enables precise age estimation and extraction of incipient and medically relevant information from PSGs that predict mortality beyond the capabilities of basic sleep metrics derived from sleep staging and apnea scoring. Subjects' ages were estimated with an MAE of  $5.8 \pm 1.16$  years with model (e, Ensemble-Avg.), while basic metrics had a MAE of  $14.6 \pm 5.91$  years. We addressed the interpretability problem of deep learning methods using gradient SHAP, which suggested that the model's estimates were largely driven by clinically known waveforms (e.g., sleep-stage transitions and apnea). We found that 10-year increments in AEE of the (e, Ensemble-Avg.) model was associated with increased all-cause mortality rate of 29% (HR = 1.29, 95% CI: 1.20–1.39) and increased cardiovascular mortality rate of 40% (HR = 1.40, 95% CI: 1.21–1.62). For a 60-year-old subject, the difference of –10 and +10 years in AEE translates to a decreased life expectancy of 8.7 years (CI: 6.1–11.4) for Cox Model 3, which adjusts for basic sleep metrics that are associated with early mortality.

The AE models performed well on the test set and generalized well to the HomePAP study test set with a MAE of  $8.16 \pm 3.75$  years

**Table 2.** Mortality hazard ratios per 10-year increment in AEE in the combined data of the Sleep Heart Health Study, the Wisconsin Sleep Cohort, and the MrOS Sleep Study.

		Cox Model 1 HR (95% CI)	Cox Model 2 HR (95% CI)	Cox Model 3 HR (95% CI)
All-cause	(a) Central EEG	1.12 (1.07–1.17)	1.15 (1.10–1.20)	1.11 (1.06–1.16)
	(b) EEG+EOG+EMG	1.11 (1.06–1.17)	1.17 (1.11–1.24)	1.14 (1.08–1.20)
	(c) ECG	1.08 (1.04–1.12)	1.09 (1.06–1.13)	1.07 (1.03–1.11)
	(d) Respiratory	1.04 (1.00–1.09)	1.10 (1.04–1.16)	1.09 (1.03–1.15)
	(e) Ensemble–Avg.	1.23 (1.15–1.31)	1.38 (1.28–1.49)	1.29 (1.20–1.39)
Cardiovascular	(a) Central EEG	1.21 (1.11–1.32)	1.24 (1.14–1.36)	1.17 (1.07–1.28)
	(b) EEG+EOG+EMG	1.13 (1.03–1.25)	1.21 (1.09–1.34)	1.15 (1.04–1.28)
	(c) ECG	1.15 (1.08–1.22)	1.16 (1.09–1.24)	1.11 (1.04–1.19)
	(d) Respiratory	1.04 (0.95–1.13)	1.09 (0.97–1.22)	1.07 (0.96–1.19)
	(e) Ensemble–Avg.	1.36 (1.20–1.54)	1.58 (1.37–1.83)	1.40 (1.21–1.62)

The mortality analysis was performed with ( $n = 9386$ , deaths = 3045) for all-cause mortality and ( $n = 9188$ , death = 976) for cardiovascular mortality. *HR* hazard ratio, *AEE* age estimate error. Model 1: age. Model 2: age, sex, body mass index, race, smoking status, education level, daily alcohol intake, daily caffeine intake, benzodiazepines, sedatives, antidepressants, and cohort. Model 3: Model 2 + wake after sleep onset, N2%, REM%, arousal index, apnea-hypopnea index, sleep time with blood oxygen saturation below 80%, Epworth Sleepiness Scale Score, hypertension, congestive heart failure, history of heart attack, stroke, and type 2 diabetes.



**Fig. 4** Survival curve for all-cause mortality with an AEE varying  $\pm 10$  years. The survival curve was generated for all data ( $n = 9386$ , deaths = 3045) and model (e, Ensemble–Avg.) using the Cox proportional hazards model 3 shown in Supplementary Table 3. The 95% CI express the uncertainty in the modeled hazard ratio. AEE age estimate error, CI confidence interval.

for the (e, Ensemble–Avg.) model. Based on these results, we expect the model to generalize to new data recorded in adult subjects aged 20–90 from other clinics, obtaining MAEs between 5.8 and 8.16 years. Calibrating the AE in new, unseen populations may however be necessary to achieve a MAE of 5.8 years in these instances. Further, although the model was trained using data that included children, this data was limited in amount and age range, so our model is not validated for use in children. We however note that validating similar age estimates in children in a separate study could have great interest for the study of neurodevelopmental disorders in children.

We found that model (e, Ensemble–Avg.) was biased for older subjects, estimating preferentially a younger age. This may be caused by either a regression to the mean of the predictions or by unhealthy subjects having died in the older (>80 years) population, i.e., a type of survival bias. Regression to the mean is a difficult issue to handle in non-linear models. The AE models output layer did not have any non-linear activation function, however, the AE still seem to have a non-linear clipping of AEs (e.g., around 83 years in Fig. 1). Given that the models have an uncertainty, it will drive the estimate

away from the edge case of 90 years, on average this estimate would increase the loss. A similar effect is observed for young subjects in both the test set and HomePAP set, which exhibit systematic overestimation. It is likely that adjusting for this bias observed in the test set would improve performance in new data. Moreover, model (e, Ensemble–Avg.) had a significant ( $p = 2 \cdot 10^{-9}$ ) night-to-night variability, which may result from of the first-night effect. However, more PSGs with multiple nights are necessary to confirm this. Using multiple PSGs for age estimation may alleviate this problem.

A previous study used a linear model of sleep staging features based on EEG only to model “brain age” and reported a MAE of 7.6 years<sup>27</sup>. However, the results are difficult to compare as the dataset, age ranges, and investigated PSG signals differ.

The deep learning AE models appeared to largely rely on patterns that are known to be related to aging such as sleep fragmentation<sup>24</sup>. The relevance attribution analyses showed that transitions to deeper sleep would cause model (b; EEG + EOG + EMG) to estimate a lower age. The analysis of arousal and sleep apnea (Supplementary Fig. 2) showed that these modulate the AE. Relevance scores were computed using a baseline of zero, which affects how the relevance scores should be interpreted. For example, relevance scores were increased after but not during apnea/hypopnea events; however, this is expected as low amplitude breaths are likely healthier than the baseline of zero amplitude in complete apnea. Moreover, the gradient SHAP method assumes an independent and linear attribution from each sample to the AE<sup>41,42</sup>, which is not capable of accurately describing PSG patterns or the processing in the deep neural network. Therefore, we can only argue that the models probably use non-linear statistics related to these known patterns without strictly summarizing sleep patterns to the frequency of binarized events. Alternatively, we could have interpreted the model attention network weights, however, the long short-term memory networks render these weights difficult to interpret.

Survival analysis found that greater AEE was associated with increased all-cause and cardiovascular mortality. In the Cox Models for all AE models, AEEs had larger hazard ratios while controlling for demographics and medication than controlling for only age or including health and basic sleep metrics. We infer from this that (1) the AE is more meaningful when knowing demographics and medication, and (2) the AEE is not fully explained by basic sleep metrics such as sleep-stage distributions, ArI, and AHI. It is thus evident that a PSG contains much more information than what is

summarized in basic sleep metrics. Our analysis (see Supplementary Table 5) of AEE in relation with morbidities found associations to T2D, hypertension, CHF, and history of heart attack, but pathways underlying these associations are difficult to identify. Short sleep duration in insomnia has been shown to be associated with T2D<sup>43</sup>, which may explain the association in model (a, central EEG), and (b; EEG + EOG + EMG). Model (c, ECG) was associated with hypertension, CHF, and history of heart attack, which was expected as these factors affect the morphology of ECG. Hypertension was associated with increased AEE in all but model (d, Respiratory). A sensitivity analysis that excluded subjects with hypertension (see Supplementary Table 13) showed that AEE was still associated with increased mortality, although the effects were smaller. In future studies, the association between AEE and mortality risk should be investigated in completely unseen cohorts to study the generalizability of this effect.

A strength of this study is the inclusion of multiple cohorts, likely increasing generalizability of our models. This is however also a limitation as measuring sleep with a common instrumentation and in a more controlled environment could have better predictive power by reducing technical noise, first-night effect, variation in recording equipment, electrode placement, room temperature, and external noise, etc. Another limitation is that sleep varies from night-to-night and our AE relies on only one sleep study per subject. It is likely that multiple examinations per subject and establishing trajectories of aging would have stronger predictive power.

Other approaches for age estimation have relied on epigenetics<sup>44</sup>, proteomics<sup>45</sup>, neuroimaging<sup>46,47</sup>, etc., but few of these markers have been linked to hard outcomes such as mortality. Freire-Aradas et al. found that 7 DNA methylation markers estimated age with a median age error of  $\pm 3.07$  years<sup>44</sup>. A systematic review of proteomic studies found that a 83-protein could estimate age with a MAE of 5.5 years<sup>45</sup>. Cole et al. leveraged T1-weighted magnetic resonance imaging (MRI) to estimate age with a MAE of 5.02 years. Moreover, each 1-year increment in this AE was associated with a 6.1% increased relative risk of all-cause mortality<sup>47</sup>. This corresponds to a hazard ratio of 1.29 for a 10-year increment, which is close to the hazard ratio we report in this study for the ensemble model. Advantages of PSG over these methods include being non-invasive, less expensive, and more accessible. It is also notable that sleep, should causality be demonstrated in future studies, can be modified by well-established behavioral and pharmacological therapies, unlike many of these other proxies.

In that sense, the AE may also serve as an outcome measure in adult subjects (20–90 years) for interventions in both clinical and research settings. Moreover, the AE could potentially serve as an easily understood marker of health for patients and the general public. In contrast, current sleep quality measures such as SE and N3% can be difficult to interpret for a given sleep clinic patient. Thereby, a sleep-based AEE could improve health literacy<sup>48</sup> among patients. A recent Danish study found that interviews based on body age assessments motivates health promotion in the workplace, which lead to a decrease in smoking and metabolic syndrome among the employees<sup>49</sup>. Moreover, a meta-analysis found that health literacy was correlated with treatment adherence, especially among vulnerable groups<sup>50</sup>. These findings may apply to sleep health as well, which could lead to better adherence to treatment such as lifestyle changes and continuous positive airway pressure (CPAP) therapy. This could be interesting to investigate in future studies.

Finally, our predictions of mortality are estimated through cardiovascular, respiratory, and brain activity related to aging, which we hypothesize is a likely proxy of premature aging, but not likely the sole or even main predictor of mortality. This is illustrated by the fact we recently found that reduced REM sleep amounts also significantly predicted mortality in these same samples<sup>13</sup>, and as shown in Cox Model 3 in Supplementary Table 3 that adjust for REM sleep%, addition of REM sleep to the Cox

Model did not diminish predictive effect of AEE on mortality. Clearly, other factors than AEE in the PSG are likely incipient biomarkers of poor health predicting mortality or new-onset morbidity. Additional approaches aiming at directly predicting mortality<sup>51</sup> and the development of cardiovascular and brain morbidity in these cohorts with and without controlling for AEE may help to uncover additional information in PSG recordings.

## METHODS

### Data description

Diversity of data is a necessity for the success of a supervised deep-learning algorithm<sup>52</sup>. Olesen et al. showed that both data quantity and diversity were essential for automatic sleep staging, a supervised learning task, using polysomnography data<sup>53</sup>. Diversity of data can be ensured using polysomnography recordings from multiple study cohorts with different study objectives and patient populations.

In this study, we included participants with a wide age range from seven study cohorts: STAGES<sup>31,32</sup>, the SSC<sup>33,34</sup>, the WSC<sup>4,34</sup>, the SHHS<sup>35,36</sup>, the MrOS<sup>36–38</sup>, the CFS<sup>36,39</sup>, and HomePAP Study<sup>36,40</sup>. Access to the SHHS, MrOS, CFS, and HomePAP Study was granted through the National Sleep Research Resource<sup>36</sup>. This study was approved by institutional review boards and written informed consent was obtained from all participants. The included study cohorts are briefly described in the subsections below:

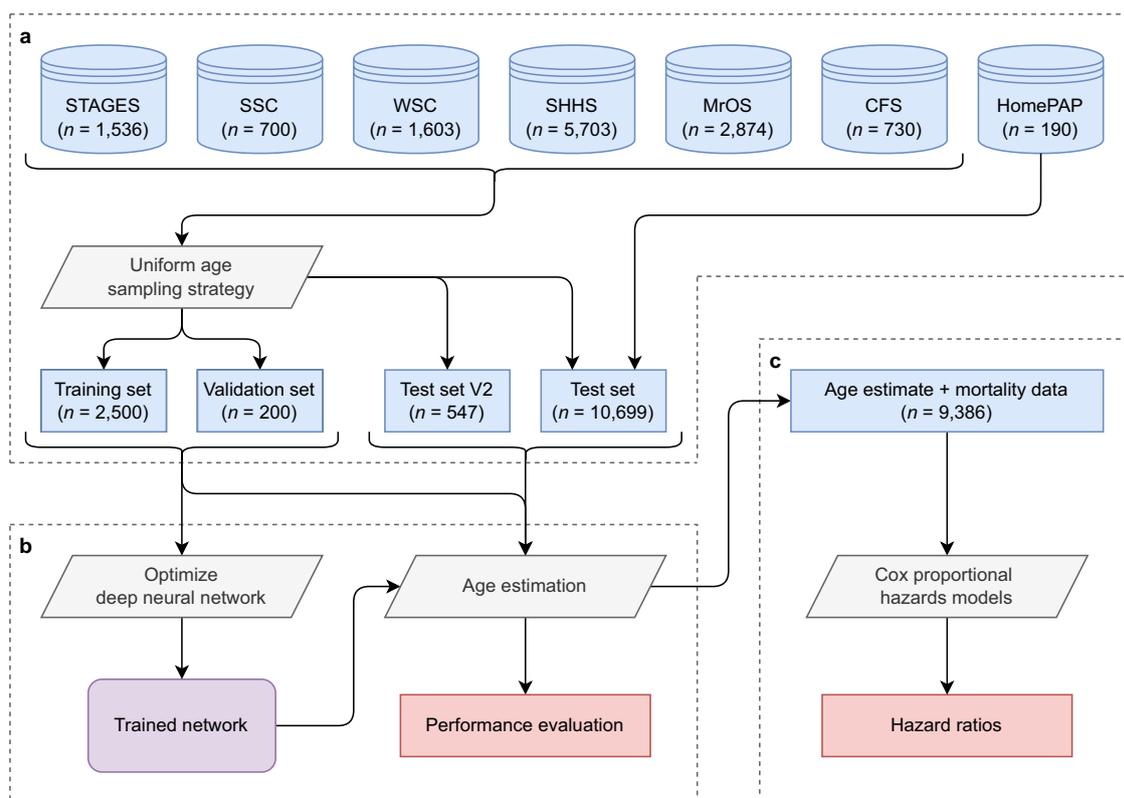
*The stanford technology analytics and genomics of sleep.* The STAGES<sup>31,32</sup> is a prospective cross-sectional multi-site cohort designed to investigate the relationship between different sleep-related data including in-lab polysomnography, questionnaire data, genomics, actigraphy data etc. A total of 1859 PSGs were recorded in 1627 participants of ages between 13 and 83 at the following 6 clinical sites: Stanford University, Bogan Sleep Consultants, Geisinger Health, Mayo Clinic, MedSleep, and St. Luke's Hospital. A total of 1536 PSGs in 1494 participants were included, while the remaining PSGs were excluded for being a split-night study or due to missing annotations. The study was approved by institutional review boards at each site.

*The Wisconsin sleep cohort.* The WSC<sup>4,34</sup> is an ongoing longitudinal population-based cohort of employees from Wisconsin state agencies, and it approximates a population-based sample, although they are generally more overweight<sup>4</sup>. A total of 1682 PSGs in 962 participants was included, which aged between 37 and 78. The participants were tracked through 2018 and deaths were identified by matching social security numbers with death record sources<sup>13</sup>. A detailed description of the cohort can be found in Young et al.<sup>4</sup> and Moore et al.<sup>34</sup>. Cardiovascular mortality was categorized using the same rules as Leary et al.<sup>13</sup>. The study has been reviewed and approved by the University of Wisconsin Institutional Review Board.

*The Stanford sleep cohort.* The SSC<sup>33,34</sup> is a cohort of patients who underwent in-lab PSG at the Stanford Sleep Clinic. A total of 700 independent PSGs was included in patients aged between 13 and 90. A detailed description of the cohort can be found in Andlauer et al.<sup>33</sup>.

*The MrOS sleep study.* The MrOS Sleep Study<sup>36–38</sup> is a multi-site cohort of older men to study the association between sleep disorders and vascular disease, falls, fractures, and mortality. A total of 2874 male participants were included who underwent full in-home PSG recording. Vital status was determined based on contact every 4 months, or in case of no response, by their next-of-kin. Reported deaths were confirmed by centralized review of death certificates<sup>13,37</sup>. Deaths through August 2018 were included in these analyses. Cardiovascular mortality was categorized using the same rules as Leary et al.<sup>13</sup>. The study was approved by the institutional review board at each of the six sites: University of Alabama at Birmingham, University of Minnesota, Stanford University, University of Pittsburgh, Oregon Health and Science University, and University of California, San Diego.

*The Cleveland family study.* The CFS<sup>36,39</sup> is a large family-based study of sleep apnea, consisting of probands with sleep apnea, neighborhood controls, and their family members. We included PSG recordings obtained in the Clinical Research Center from 730 participants of age between 6 and 88 years. The study was approved by the institutional review committee at the University Hospitals Case Medical Center.



**Fig. 5 Use of data for age estimation and evaluating mortality risk.** **a** The data from six cohorts are sampled to generate a training and validation set with a uniform age distribution. The remaining data comprises a test set, some of which has additional visits (test set V2). **b** Age estimation models are optimized and evaluated in all data. **c** Associations between increased age estimate errors and mortality risk are evaluated in all available data using Cox proportional hazards models.

**The sleep heart health study.** The SHHS<sup>35,36</sup> is a large multi-center cohort designed to study the association between sleep apnea and cardiovascular disease. We included 5703 participants, aged between 40 and 90, were studied with in-home PSG. Participants vital status was continually identified and confirmed using interviews, written annual questionnaires, contact to next-of-kin, hospital records, obituaries, and linkage with the Social Security Administration Death Master File<sup>35,54</sup>. Cardiovascular mortality was categorized as recorded by parent studies<sup>35,54</sup>. The study was approved by institutional review boards at each of the six sites: University of Arizona, Boston University, University of California-Davis, Johns Hopkins University, University of Minnesota, and New York University.

**The home positive airway pressure study.** The HomePAP Study<sup>36,40</sup> is a multi-site randomized controlled trial with the aim of comparison in-lab PSG and in-home unattended portable monitoring for diagnosis of obstructive sleep apnea and CPAP treatment. We included 190 patients of age between 20 and 80 with in-lab PSG without full or split-night CPAP at one of 7 sites: Case Western Reserve University affiliates (University Hospitals, MetroHealth Medical Center, and Cleveland Clinic), Northwestern University, University of Wisconsin in Madison, University of Minnesota, and University of Washington. The study was approved by institutional review boards at each site.

**Data use and study design.** Across the cohorts, PSGs were excluded if the participant's age was unknown, the recording was a CPAP split-night, the recording included <3 h of sleep, or if more than one of the PSG signals were missing.

To facilitate the development and testing of the AE models, the combined data were split into a training set ( $n = 2500$ ), a validation set ( $n = 200$ ), a test set ( $n = 10,699$ ), and a second test set comprised of repeat visits ( $n = 547$ ). The AE models are developed using the training and validation set, which should include diverse data at all ages. To ensure this, we propose a sampling strategy with uniform age distribution in favor of the commonly used random sampling. Firstly, patients who used CPAP or

had any known neurological disorders including narcolepsy and RBD were allocated to the test set. Data was sampled for the training set by iteratively excluding data with the most represented age, cohort, and sex, see Supplementary Table 16 for details. Similarly, the validation set uses the same algorithm with the remaining data. The high-level flow of data from each cohort to various sets used for age estimation and evaluation of mortality risk is shown in Fig. 5.

The test set used the remaining data, which was not uniform but can be analyzed stratified by age. Participants with an age >89 were recorded as being 90, therefore we chose to exclude these from the test set. Moreover, data from the HomePAP Study ( $n = 190$ ) was left out of the remaining test set for an additional test set, which provided an unbiased performance estimate as no data from the cohort is included in the training or validation set. Supplementary Fig. 3 shows the distribution of age and cohorts across the training, validation, and test sets. Supplementary Table 17 shows the distribution of basic PSG metrics across the included cohorts. The apneas and hypopneas were scored in agreement with AASM guidelines<sup>1</sup>, which requires associations with either a 3% desaturation or an arousal for hypopneas. Arousals were either scored manually in agreement with the AASM guidelines<sup>1</sup> (CFS, MrOS, SHHS, HomePAP) or automatically scored using a previously validated method<sup>21</sup> (STAGES, WSC, SSC) when manual annotations were missing.

### Preprocessing of polysomnographic signals

The PSG data included in this study have been recorded at many clinical sites with varying signal montages, environments, technicians, equipment, software, and acquisition settings. These differences are addressed in the preprocessing step to both standardize the data and eliminate signal artifacts. Specifically, we implemented a preprocessing that can (1) select the appropriate signal derivations; (2) resample signals to a desired and standardized sampling frequency; (3) eliminate signal artifacts; and (4) normalize signal amplitudes.

A PSG recording involves measuring many physiological signals and these can vary between recordings. Most commonly, the PSG recording

includes electroencephalography (EEG) signals, electrooculography (EOG) signals (left and right), electromyography (chin), electrocardiography (ECG), nasal pressure, oral airflow, plethysmography belts (abdominal and chest), and blood oxygen saturation. Except for frontal and occipital EEG, we included all of these as they were available in almost all cohorts. Potential missing signals was substituted for flat signals of zeros.

The convolution neural networks (CNNs) assume a constant sampling rate, therefore, the signals are resampled to a sampling frequency of 128 Hz, which enables all signals to be stacked in one tensor. This frequency was chosen as it preserves most relevant information while still imposing a relatively low computational burden. The resampling was implemented with a finite impulse response (FIR) low-pass filter with a Kaiser window. However, the blood oxygen saturation was resampled with linear interpolation.

Thereafter, signals were filtered using infinite impulse response (IIR) filters to eliminate artifacts and ensure that signals contained similar spectral content across recordings. The IIR filters were implemented as elliptic filters with an order of 16, a maximum passband ripple of 1 dB, and minimum stopband attenuation of 40 dB. The cut-off frequencies for the filters were the following: EEG and EOG: band-pass (0.3–45 Hz); EMG: high-pass (10 Hz); ECG: high-pass (0.3 Hz); nasal pressure: high-pass (0.1 Hz); airflow and plethysmography belts: band-pass (0.1–15 Hz); and blood oxygen saturation: no filtering. All filters were applied forwards and backwards to avoid signal phase distortion.

Finally, the signal amplitudes, except for the blood oxygen saturation, were normalized such that  $-1$  and  $1$  corresponded to the 5th and 95th percentiles. Although, the blood oxygen saturation was normalized such that  $-1$  and  $1$  corresponded to 60% and 100% saturation. The normalization enables efficient training of neural networks<sup>52</sup>.

## Deep learning framework for age estimation models

The proposed deep learning framework for AE was designed to input  $C$  number of preprocessed PSG signals of  $T$  samples  $\mathbf{x} \in \mathbb{R}^{C \times T}$  and output an estimate of the subject's age  $\hat{y} \in \mathbb{R}_+$ . In the subsections below, the network architecture, the optimization approach, performance testing, and the interpretation of models is presented.

A challenge in end-to-end deep learning processing of PSG recordings is the huge data size, which usually is of 8 h corresponding to an input dimension of  $\mathbf{x} \in \mathbb{R}^{12 \times (128 \times 60 \times 60 \times 8)}$  corresponding to roughly 177 MB in 32-bit float. Directly optimizing a network to map the whole night's PSG to estimate age is practically infeasible as intermediate network activations must be saved for optimization through backpropagation. Therefore, we chose to optimize the networks in two phases.

Phase (1): Estimating age based on 5-min epochs of PSG data  $\mathbf{x}_i \in \mathbb{R}^{C \times (128 \times 60 \times 5)}$ .

Phase (2): Estimating age based on the latent space learned  $\mathbf{z}_i \in \mathbb{R}^M$  in phase 1 (at the network layer preceding the output layer) for all 5-min epochs of length  $q$  in a whole night's recording  $\mathbf{z} \in \mathbb{R}^{M \times \lfloor T/q \rfloor}$ .

Thereby, the networks first learn signal patterns in 5-min epochs that are associated with aging, and secondly, distributions of these patterns across the night taken into consideration by the networks.

**Neural network architecture of age estimation models.** The network incorporates a series of structures that have shown success in sleep-stage classification from PSG data<sup>19,53,55,56</sup>, image classification<sup>57</sup>, and natural language processing<sup>58</sup>.

As shown in phase (1) in Supplementary Fig. 4, 5-min epochs of data  $\mathbf{x}_i$  are processed through a channel mixing layer, a CNN using inverted residual bottleneck blocks (see Supplementary Fig. 5), a bi-directional long short-term memory<sup>59</sup> (Bi-LSTM) layer, an additive attention<sup>58,60</sup> layer, and two dense layers, which produces an estimate of age  $\hat{y}_{P1}$ .

As shown in phase (2) in Supplementary Fig. 4, the latent space  $\mathbf{z}_i$  is concatenated from the layer activation in phase (1) at the last layer and the average activation after the Bi-LSTM layer to summarize the 5-min epochs of data  $\mathbf{x}_i$ . Like phase (1), the whole night's latent space  $\mathbf{z}$  is processed through a Bi-LSTM layer, an additive attention layer, and two dense layers, which produce a final AE  $\hat{y}_{P2}$ .

The implementation details of each neural network type<sup>61–65</sup> are presented in the Supplementary Notes.

**Optimization scheme for age estimation models.** The network was optimized in two phases as outlined in Supplementary Fig. 4 to both lower the computational burden and increase the amount of training data. The Huber loss used as the objective function to minimize and was defined

as

$$L_H = f(x) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{for } |y - \hat{y}| < 5 \\ 5(|y - \hat{y}| - \frac{1}{2}5), & \text{otherwise,} \end{cases} \quad (1)$$

which corresponds to an L2 loss for an error  $<5$  years and L1 loss otherwise. This loss weighs outliers less than an L2 loss while retaining a continuous gradient. The loss was further divided by a factor of 112.5 such that an error of 25 years corresponds to a loss of 1. Using this loss and L2 weight decay (not counting network bias's), the network was optimized using Adam optimization<sup>66</sup> with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$ .

Additional optimization settings and hyperparameter tuning methods<sup>67</sup> are described in Supplementary Notes.

We experimented with various combination of PSG signals: (a, Central EEG) C3-A2, C4-A1; (b, EEG+EOG+EMG) C3-A2, C4-A1, L-EOG, R-EOG, chin EMG; (c, ECG) ECG; (d, Respiratory) airflow, nasal pressure, thoracic and abdominal belts, SaO2.

Moreover, an ensemble model (e, Ensemble-Avg.) was developed based on models (a), (b), (c), and (d).

Finally, as a comparison to basic sleep summary measures, a linear regression model using sex, BMI, ArI, AHl, TST, WASO, and percentage of N1, N2, N3, and REM sleep was developed.

**Performance quantification of age estimation models.** The performance of the AE was quantified using mean absolute error (MAE) and Pearson's correlation coefficient. The test set was not characterized by a uniform age distribution; therefore, we measured the MAE stratified by 5-year age intervals (MAE<sub>i</sub>) with intervals ([20, 25], [25, 30], ..., [85–89]). The average MAE across age intervals MAE<sub>i</sub> was used as a final measure of performance.

**Interpretation of age estimation models.** Deep neural networks are traditionally considered black boxes due to their complexity, which is of high concern generally and even more so in a clinical setting. However, in recent years several methods have been proposed that can interpret network decisions in a meaningful way<sup>68,69</sup>. We applied gradient SHAP<sup>41,42</sup> to distribute relevance scores to each PSG sample using phase (1) of the optimized networks. To remove noise, the sample relevance scores were filtered by a Gaussian window with a length of 10 s and standard deviation of 0.234 s. Relevance scores were averaged around transitions of manually scored hypnograms, arousal, and apnea/hypopnea events in PSGs from the CFS, MrOS, and SHHS cohort in the test set. Here we expected to see increases in relevance scores arousals, transitions to lighter sleep, and sleep apnea.

Moreover, we examined statistical associations between the AEE and conventional sleep parameters from manual scoring.

## Association between age estimation and mortality

The usefulness of the AEE as a marker for sleep health was examined by studying its association with all-cause mortality. This analysis was performed in the SHHS, MrOS, and WSC.

**Statistical analyses.** We considered that missing data were missing at random and these were imputed using multivariate imputation by chained operationalized equations using R 4.0.4 MICE package<sup>70</sup>. Information about CPAP had a lot of missing data (0 for WSC, 5603 for SHHS, and 2671 for MrOS), therefore, we excluded the few subjects ( $n = 74$ ) that used CPAP from these analyses.

We employed Cox proportional hazards models to evaluate associations between AEE and all-cause mortality. The results are reported as hazard ratios (HR) along with their 95% confidence intervals (CI) for every 10 years increase in AEE, which is close to the standard deviation of AEE.

The Cox proportional hazards models controlled for a combination of variables based on clinical and empirical knowledge<sup>13</sup>. Covariates were included in three combinations to investigate if the association was dependent on these covariates. Cox model 1 adjusted for age; Cox model 2 that included covariates we clinically known or suspect to influence mortality: age, sex, BMI, race, education, smoking status, daily alcohol intake, daily caffeine intake, medication use (antidepressants, benzodiazepines, and sedatives), and study site. Cox model 3 that included covariates from Cox model 2 and covariates empirically found to affect mortality in the MrOS cohort using 6-fold cross validation<sup>13</sup>: NREM%, REM%, SaO2-80, WASO, ArI, ESS, congestive heart failure, chronic obstructive pulmonary disease, type 2 diabetes, heart attack, and stroke. The proportional hazards assumption for AEE was confirmed graphically by analyzing the scaled Schoenfeld residuals.

A summary of all covariates was computed for each quartile of the AEE corrected for age variation (AEEc) for model (h, Ensemble-Avg. EEG). The AEEc was computed as the residuals in the linear regression model  $AEE = 1 + \text{age} + \varepsilon$ , i.e., where  $\varepsilon$  is the AEEc.

Fitted Cox proportional hazards models were formulated as survival functions  $S_0^{KM}(t)^{z(x)}$ , where  $S_0^{KM}(t)$  is the baseline survival function and  $z(x) = \exp(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)$ . The survival functions were plotted with the AEE as  $\pm 10$  years, corresponding to the estimated hazard ratio  $\exp(\beta_{AEE} \times 10)$ . Moreover, similar to a previous approach<sup>28</sup>, we computed the effect of an increased AEE on life expectancy by extending the survival curve and computing the difference in curve area. The baseline survival curves were extended by fitting a Weibull distribution  $S_0^W(t)$ . Life expectancy was computed as the area of  $S_0^W(t)^{z(x)}$  with age set to 40, 60, or 80 years and the other covariates to their median in that age range  $\pm 10$  years. The difference in life expectancy was found by subtracting the LE for  $AEE = -10$  and  $AEE = 10$ .

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## DATA AVAILABILITY

Polysomnography data included in this study was subject to data sharing agreement but is available upon reasonable request from E.M. (for STAGES and SSC), P.E.P. (WSC), or upon request from the NSRR<sup>36</sup> (SHHS, MrOS, CFS, and HomePAP).

## CODE AVAILABILITY

The AE models were implemented, trained, and tested using PyTorch<sup>71</sup> v. 1.7.1. All source code is available at <https://github.com/abrinkk/psg-age-estimation>, which includes instructions for use.

Received: 10 January 2022; Accepted: 10 June 2022;

Published online: 22 July 2022

## REFERENCES

- Berry, R. et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications. Version 2.* (American Academy of Sleep Medicine, 2018).
- Magalang, U. J. et al. Agreement in the scoring of respiratory events and sleep among international sleep centers. *Sleep* **36**, 591–596 (2013).
- Koch, H. et al. Breathing disturbances without hypoxia are associated with objective sleepiness in sleep apnea. *Sleep*. <https://doi.org/10.1093/sleep/zsx152> (2017).
- Young, T. et al. Burden of sleep apnea: rationale, design, and major findings of the Wisconsin sleep cohort study. *Wis. Med. J.* **108**, 246–249 (2009).
- Young, T. et al. The occurrence of sleep-disordered breathing among middle-aged adults. *N. Engl. J. Med.* **328**, 1230–1235 (1993).
- Nieto, F. J. Association of sleep-disordered breathing, sleep apnea, and hypertension in a large community-based study. *JAMA* **283**, 1829 (2000).
- Azarbarzin, A. et al. The hypoxic burden of sleep apnoea predicts cardiovascular disease-related mortality: the osteoporotic fractures in men study and the sleep heart health study. *Eur. Heart J.* **40**, 1149–1157 (2019).
- Aritake, S. et al. Prevalence and associations of respiratory-related leg movements: the MrOS sleep study. *Sleep. Med.* **16**, 1236–1244 (2015).
- Redline, S. et al. Obstructive sleep apnea-hypopnea and incident stroke. *Am. J. Respir. Crit. Care Med.* **182**, 269–277 (2010).
- Jones, S. Sleep disordered breathing and mortality: eighteen-year follow-up of the wisconsin sleep cohort. *Yearb. Pulm. Dis.* **2009**, 291–292 (2009).
- Shahrbabaki, S. S., Linz, D., Hartmann, S., Redline, S. & Baumert, M. Sleep arousal burden is associated with long-term all-cause and cardiovascular mortality in 8001 community-dwelling older men and women. *Eur. Heart J.* **42**, 2088–2099 (2021).
- Wallace, M. L. et al. Physiological sleep measures predict time to 15-year mortality in community adults: Application of a novel machine learning framework. *J. Sleep Res.* <https://doi.org/10.1111/jsr.13386> (2021).
- Leary, E. B. et al. Association of rapid eye movement sleep with mortality in middle-aged and older adults. *JAMA Neurol.* **77**, 1241 (2020).
- Yan, B. et al. Sleep fragmentation and incidence of congestive heart failure: the Sleep Heart Health Study. *J. Clin. Sleep Med.* <https://doi.org/10.5664/jcsm.9270> (2021).
- Yan, B. et al. Objective sleep efficiency predicts cardiovascular disease in a community population: the sleep heart health study. *J. Am. Heart Assoc.* **10**, 16201 (2021).
- Schenck, C. H., Boeve, B. F. & Mahowald, M. W. Delayed emergence of a parkinsonian disorder or dementia in 81% of older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder: a 16-year update on a previously reported series. *Sleep. Med.* **14**, 744–748 (2013).
- Dauvilliers, Y. et al. REM sleep behaviour disorder. *Nat. Rev. Dis. Prim.* **4**, 19 (2018).
- Högl, B., Santamaria, J., Iranzo, A. & Stefani, A. Precision medicine in rapid eye movement sleep behavior disorder. *Sleep. Med. Clin.* **14**, 351–362 (2019).
- Stephansen, J. B. et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat. Commun.* **9**, 5229 (2018).
- Perslev, M. et al. U-Sleep: resilient high-frequency sleep staging. *npj Digit. Med.* **4**, 72 (2021).
- Brink-Kjaer, A. et al. Automatic detection of cortical arousals in sleep and their contribution to daytime sleepiness. *Clin. Neurophysiol.* **131**, 1187–1203 (2020).
- Mander, B. A., Winer, J. R. & Walker, M. P. Sleep and human aging. *Neuron* **94**, 19–36 (2017).
- Li, J., Vitiello, M. V. & Gooneratne, N. S. Sleep in normal aging. *Sleep. Med. Clin.* **13**, 1–11 (2018).
- Boselli, M., Parrino, L., Smerieri, A. & Terzano, M. G. Effect of age on EEG arousals in normal sleep. *Sleep* **21**, 351–357 (1998).
- Crowley, K. The effects of normal aging on sleep spindle and K-complex production. *Clin. Neurophysiol.* **113**, 1615–1622 (2002).
- Floyd, J. A., Janisse, J. J., Jenuwine, E. S. & Ager, J. W. Changes in REM-sleep percentage over the adult lifespan. *Sleep* **30**, 829–836 (2007).
- Sun, H. et al. Brain age from the electroencephalogram of sleep. *Neurobiol. Aging* **74**, 112–120 (2019).
- Paixao, L. et al. Excess brain age in the sleep electroencephalogram predicts reduced life expectancy. *Neurobiol. Aging* **88**, 150–155 (2020).
- Ye, E. et al. Association of sleep electroencephalography-based brain age index with dementia. *JAMA Netw. Open* **3**, e2017357 (2020).
- Leone, M. J. et al. HIV increases sleep-based brain age despite antiretroviral therapy. *Sleep*. <https://doi.org/10.1093/sleep/zsab058> (2021).
- Leary, E. B. et al. 0322 Development of complex data platform for the stanford technology analytics and genomics in sleep (STAGES) study. *Sleep* **42**, A132–A132 (2019).
- Leary, E. B., Seeger-Zybok, R. K., Kushida, C. & Mignot, E. 0324 Improving our understanding of sleep by generating and sharing a large sleep cohort and data analytic tools. *Sleep* **41**, A124–A124 (2018).
- Andlauer, O. et al. Nocturnal rapid eye movement sleep latency for identifying patients with narcolepsy/hypocretin deficiency. *JAMA Neurol.* **70**, 891 (2013).
- Moore, H. et al. Design and validation of a periodic leg movement detector. *PLoS One* **9**, e114565 (2014). Penzel, T, ed.
- Quan, S. F. et al. The sleep heart health study: design, rationale, and methods. *Sleep* **20**, 1077–1085 (1997).
- Dean, D. A. et al. Scaling up scientific discovery in sleep medicine: the national sleep research resource. *Sleep* **39**, 1151–1164 (2016).
- Blackwell, T. et al. Associations between sleep architecture and sleep-disordered breathing and cognition in older community-dwelling men: the osteoporotic fractures in men sleep study. *J. Am. Geriatr. Soc.* **59**, 2217–2225 (2011).
- Orwoll, E. et al. Design and baseline characteristics of the osteoporotic fractures in men (MrOS) study—A large observational study of the determinants of fracture in older men. *Contemp. Clin. Trials* **26**, 569–585 (2005).
- Redline, S. et al. The familial aggregation of obstructive sleep apnea. *Am. J. Respir. Crit. Care Med.* **151**, 682–687 (1995).
- Rosen, C. L. et al. A multisite randomized trial of portable sleep studies and positive airway pressure autotitration versus laboratory-based polysomnography for the diagnosis and treatment of obstructive sleep apnea: The HomePAP Study. *Sleep* **35**, 757–767 (2012).
- Kokhlikyan, N. et al. Captum: a unified and generic model interpretability library for PyTorch. *arXiv*. <https://doi.org/10.48550/arXiv.2009.07896> (2020).
- Lundberg S. M. & Lee S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* <https://doi.org/10.48550/arXiv.1705.07874> (2017)
- Vgontzas, A. N. et al. Insomnia with objective short sleep duration is associated with Type 2 diabetes: a population-based study. *Diabetes Care* **32**, 1980–1985 (2009).
- Freire-Aradas, A. et al. Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system. *Forensic Sci. Int. Genet.* **24**, 65–74 (2016).
- Johnson, A. A., Shokhirev, M. N., Wyss-Coray, T. & Lehallier, B. Systematic review and analysis of human proteomics aging studies unveils a novel proteomic aging clock and identifies key processes that change with age. *Ageing Res. Rev.* **60**, 101070 (2020).
- Cole, J. H., Franke, K. & Cherbuin, N. in *Biomarkers of Human Aging*. (ed. Moskalev A) 293–328 (Springer International Publishing, 2019).

47. Cole, J. H. et al. Brain age predicts mortality. *Mol. Psychiatry* **23**, 1385–1392 (2018).
48. Liu, C. et al. What is the meaning of health literacy? A systematic review and qualitative synthesis. *Fam. Med. Community Heal* **8**, e000351 (2020).
49. Husted, K. L. S., Dandanell, S., Petersen, J., Dela, F. & Helge, J. W. The effectiveness of body age-based intervention in workplace health promotion: results of a cohort study on 9851 Danish employees. *PLoS One* **15**, e0239337 (2020). Tauler P, ed.
50. Miller, T. A. Health literacy and adherence to medical treatment in chronic and acute illness: a meta-analysis. *Patient Educ. Couns.* **99**, 1079–1086 (2016).
51. Lee, C., Zame, W. R., Yoon, J. & Van Der Schaar, M. *DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks*. Vol 32.; *Proceedings of the AAAI Conference on Artificial Intelligence* (PKP Publishing Services Network, 2018).
52. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
53. Olesen, A. N., Jørgen Jennum, P., Mignot, E. & Sorensen, H. B. D. Automatic sleep stage classification with deep residual neural networks in a mixed-cohort setting. *Sleep*. <https://doi.org/10.1093/sleep/zsaa161> (2021).
54. Punjabi, N. M. et al. Sleep-disordered breathing and mortality: a prospective cohort study. *PLoS Med.* **6**, e1000132 (2009). Patel A, ed.
55. Olesen, A. N. et al. Towards a flexible deep learning method for automatic detection of clinically relevant multi-modal events in the polysomnogram. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 556–561 (IEEE, 2019).
56. Phan, H., Andreotti, F., Cooray, N., Chen, O. Y. & De Vos, M. SeqSleepNet: End-to-end hierarchical recurrent neural network for sequence-to-sequence automatic sleep staging. *IEEE Trans. Neural Syst. Rehabil. Eng.* **27**, 400–410 (2019).
57. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen L.-C. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 4510–4520 (IEEE, California, 2018).
58. Yang, Z. et al. Hierarchical attention networks for document classification. In *Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489 (Association for Computational Linguistics, 2016).
59. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
60. Bahdanau, D., Cho, K. H. & Bengio Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings*. International Conference on Learning Representations (ICLR, 2015).
61. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proc 32nd International Conference on Machine Learning*. (PMLR, 2015).
62. Chambon, S., Thorey, V., Arnal, P. J., Mignot, E. & Gramfort, A. DOSED: A deep learning approach to detect multiple sleep micro-events in EEG signal. *J. Neurosci. Methods* **321**, 64–78 (2019).
63. Bianco, S., Cadene, R., Celona, L. & Napolitano, P. Benchmark analysis of representative deep neural network architectures. *IEEE Access* **6**, 64270–64277 (2018).
64. Terzano, M. G. et al. Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep. *Sleep. Med* **2**, 537–553 (2001).
65. Srivastava, N., Hinton, G., Krizhevsky, A. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn Res* **15**, 1929–1958 (2014).
66. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv* <https://doi.org/10.48550/arXiv.1412.6980> (2014).
67. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. & de Freitas, N. Taking the human out of the loop: a review of bayesian optimization. *Proc. IEEE* **104**, 148–175 (2016).
68. Montavon, G., Samek, W. & Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process* **73**, 1–15 (2018).
69. Kohlikiyan, N. et al. Captum: A unified and generic model interpretability library for PyTorch. *arXiv* <http://arxiv.org/abs/2009.07896> (2020).
70. Buuren, Svan & Groothuis-Oudshoorn, K. mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**, 1–67 (2011).
71. Paszke, A. et al. *31st Conference on Neural Information Processing Systems (NIPS)*, 2017).

## ACKNOWLEDGEMENTS

This research was supported by the Danish Center for Sleep Medicine, the Technical University of Denmark, and the Klarman Family Foundation. Additional support to A. B.-K. was provided by the Stibo, Oberstløjtnant Max Nørgaard & Hustru Magda Nørgaards, Otto Monsted, Augustinus, Knud Højgaard, William Demant, Vera & Carl Johan Michaelsens, Tranes, Marie & M.B. Richters Fond, and IDAs & Berg-Nielsens foundations. The Stanford Technology, Analytics and Genomics in Sleep (STAGES) study was funded by the Klarman Family Foundation. 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 U01HL64363 (Case Western Reserve University). The Cleveland Family Study (CFS) was supported by grants from the National Institutes of Health (HL46380, M01 RR00080-39, T32-HL07567, R01-46380). The Osteoporotic Fractures in Men (MrOS) Study is supported by National Institutes of Health funding. The following institutes provide support: the National Institute on Aging (NIA), the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS), the National Center for Advancing Translational Sciences (NCATS), and NIH Roadmap for Medical Research under the following grant numbers: U01 AG027810, U01 AG042124, U01 AG042139, U01 AG042140, U01 AG042143, U01 AG042145, U01 AG042168, U01 AR066160, and UL1 TR000128. The National Heart, Lung, and Blood Institute (NHLBI) provides funding for the MrOS Sleep ancillary study “Outcomes of Sleep Disorders in Older Men” under the following grant numbers: R01 HL071194, R01 HL070848, R01 HL070847, R01 HL070842, R01 HL070841, R01 HL070837, R01 HL070838, and R01 HL070839. See MrOs online public data release website: <https://mrosonline.ucsf.edu>. The Home Positive Airway Pressure study (HomePAP) was supported by the American Sleep Medicine Foundation 38-PM-07 Grant: Portable Monitoring for the Diagnosis and Management of OSA. This Wisconsin Sleep Cohort Study was supported by the U.S. National Institutes of Health, National Heart, Lung, and Blood Institute (R01HL62252), National Institute on Aging (R01AG036838, R01AG058680), and the National Center for Research Resources (1UL1RR025011). The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24 HL114473, 75N92019R002).

## AUTHOR CONTRIBUTIONS

A.B.-K. laid out the design of the study, conducted most analyses, and wrote most of the manuscript. E.B.L. assisted in the mortality analyses. H.S. and M.B.W. assisted in design of the study. K.L.S., P.E.P., P.M.C., N.E.L., and S.R. contributed datasets. H.B.D.S. and P.J. participated in the design of the study and supervised the analyses. E.M. participated in the design of the study, contributed dataset, assisted in the writing, and supervised the analyses. All authors contributed to manuscript writing and helped revising the manuscript.

## COMPETING INTERESTS

The authors declare no competing non-financial interests and but the following competing financial interest: E.B.L. is now a full-time employee of Jazz Pharmaceuticals who, in the course of this employment, has received stock options exercisable for, and other stock awards of, ordinary shares of Jazz Pharmaceuticals, plc.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41746-022-00630-9>.

**Correspondence** and requests for materials should be addressed to Andreas Brink-Kjaer or Emmanuel Mignot.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022