

<https://doi.org/10.1038/s41746-025-01607-0>

# Continuous sleep depth index annotation with deep learning yields novel digital biomarkers for sleep health

Check for updates

Songchi Zhou<sup>1</sup>, Ge Song<sup>2</sup>, Haoqi Sun<sup>3</sup>, Deyun Zhang<sup>4</sup>, Yue Leng<sup>5</sup>, M. Brandon Westover<sup>3</sup> & Shenda Hong<sup>1</sup> ✉

Traditional sleep staging categorizes sleep and wakefulness into five coarse-grained classes, overlooking subtle variations within each stage. We propose a deep learning method to annotate continuous sleep depth index (SDI) with existing discrete sleep staging labels, using polysomnography from over 10,000 recordings across four large-scale cohorts. The results showcased a strong correlation between the decrease in sleep depth index and the increase in duration of arousal. Case studies indicated that SDI captured more nuanced sleep structures than conventional sleep staging. Clustering based on the digital biomarkers extracted from the SDI identified two subtypes of sleep, where participants in the disturbed subtype had a higher prevalence of several poor health conditions and were associated with a 33% increased risk of mortality and a 38% increased risk of fatal coronary heart disease. Our study underscores the utility of SDI in revealing more detailed sleep structures and yielding novel digital biomarkers for sleep medicine.

Sleep is essential to human health, and poor sleep poses threats to people's daily life<sup>1,2</sup> and is linked to numerous diseases<sup>3,4</sup>. Sleep disorders, including sleep fragmentation and obstructive sleep apnea (OSA), are widespread and associated with adverse health outcomes<sup>5,6</sup>. In sleep medicine, sleep staging based on polysomnography (PSG) has been an indispensable part of revealing sleep structures for disease diagnosis<sup>7-9</sup>. According to the American Academy of Sleep Medicine (AASM) guidelines<sup>10</sup>, sleep and wakefulness are classified into five stages: wake (W), rapid eye movement (REM) (R), non-REM stage 1 (N1), non-REM stage 2 (N2), and non-REM stage 3 (N3) for every non-overlapping 30-second. Despite its importance, manual PSG scoring is labor-intensive and prone to variability<sup>11,12</sup>. Consequently, numerous machine-learning methods have been developed to automate sleep staging with performance comparable to human experts<sup>13-17</sup>. However, current sleep staging is too coarse to accurately reflect sleep depth due to detailed differences in sleep structures within the same stages<sup>18-20</sup>.

A continuous measure of sleep depth, reflecting the likelihood of being aroused from sleep, holds more value than discrete sleep staging results with respect to research on populations with sleep fragmentation. It has traditionally been studied using external stimuli<sup>21,22</sup>. These methods may fail to reflect instantaneous arousal because of the application of stimulus over a long period when sleep depth can already change. Electroencephalogram (EEG) delta power is commonly considered to be closely related to sleep

depth. Notably, the odds ratio product (ORP) is introduced as a continuous estimate of sleep depth, derived from the relationship between EEG power in different frequencies<sup>23</sup>. Subsequent studies have validated its effectiveness as an index of sleep depth<sup>24-28</sup>. Recent research demonstrates that principal component analysis (PCA) on pre-computed sleep-stage clusters can distinguish sleep stages in a low-dimensional sub-space<sup>29</sup>. These methods, however, rely solely on EEG data, while AASM recommends using PSG for a comprehensive analysis<sup>10</sup>. Additionally, these measures often involve manually computed features that may not capture nuanced sleep structures.

Artificial intelligence (AI), particularly deep learning<sup>30</sup>, has become increasingly popular in sleep medicine<sup>31,32</sup>. Specifically for sleep depth annotation, AI has the potential to process vast amounts of PSG data and uncover detailed information overlooked by clinicians using some off-the-shelf sleep labels (e.g., sleep staging labels and respiratory events). Given the ordinal nature of Non-Rapid Eye Movement (NREM) stages, expert-labeled staging results can help derive a measure of sleep depth. However, the existing sleep staging labels are coarse-grained as five discrete classes, prohibiting direct supervised training to generate a continuous sleep depth measure. Inspired by the idea of learning to rank widely used in various fields<sup>33-35</sup>, we may use ranking-based methods to guide the model to assign higher values to sleep epochs labeled with sleep stages conventionally regarded as deeper sleep. The stages from N1 to N3 represent progressively

<sup>1</sup>National Institute of Health Data Science, Peking University, Beijing, China. <sup>2</sup>Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai, China. <sup>3</sup>Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. <sup>4</sup>HeartVoice Medical Technology, Hefei, China. <sup>5</sup>Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, CA, USA. ✉e-mail: hongshenda@pku.edu.cn

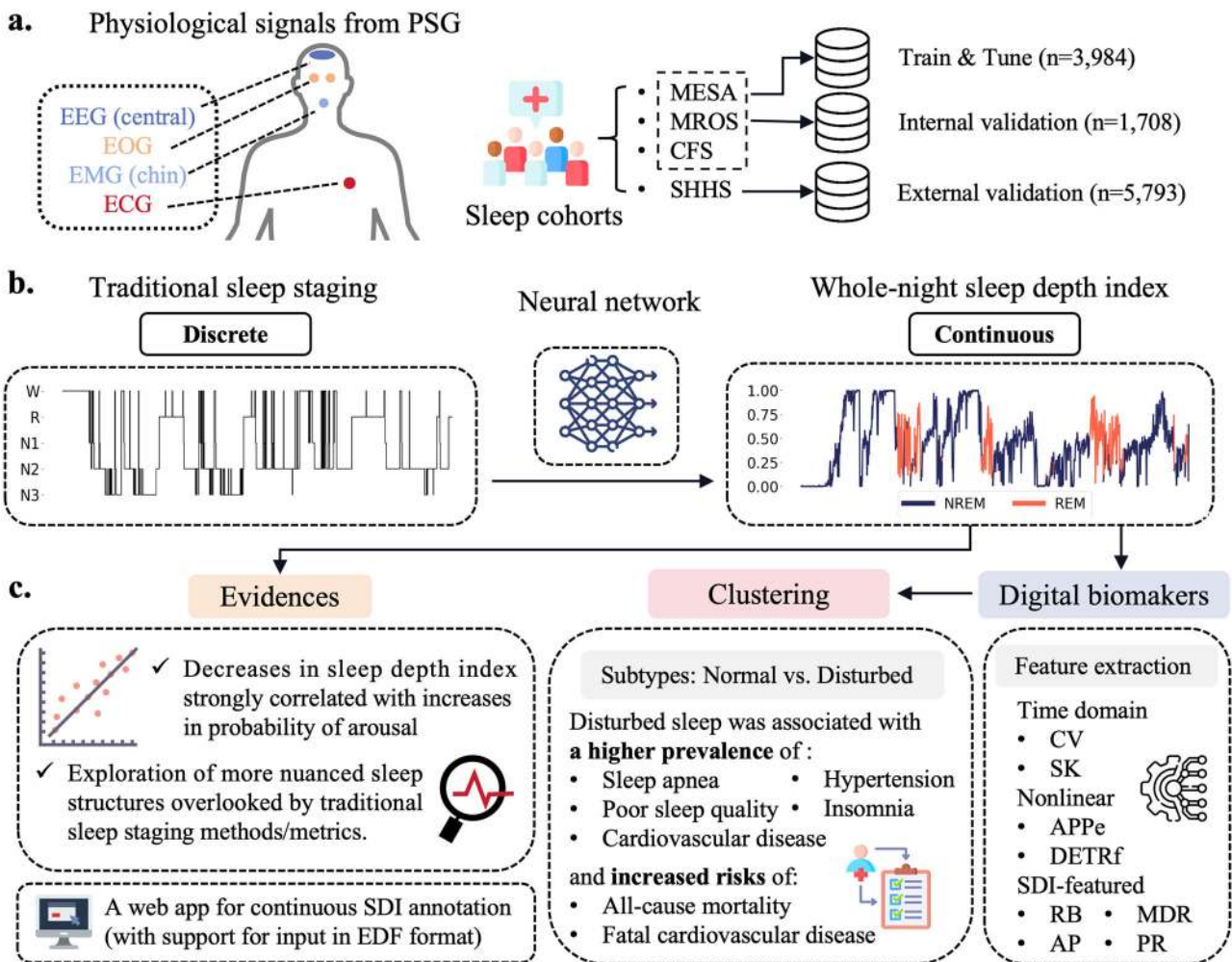
deeper sleep<sup>36</sup>, fitting well within the learning-to-rank framework. The REM stage, due to its unique significance in clinical settings<sup>37–39</sup>, requires careful handling in comparisons with NREM stages. Nevertheless, REM is deeper than wakefulness and can be integrated into the ranking process, though its comparison with other NREM stages needs caution. Even though the acquirement of the whole-night sleep depth index can serve as a valuable tool to help clinicians inspect the sleep structure from a more detailed perspective, there is no guideline on how to use this whole-night sleep depth index for routine clinical practice for sleep health. Since the whole-night sleep depth index is a type of time series, we can extract several features such as basic time-domain features or complexity-related features as novel digital biomarkers, and investigate their indications for various health conditions.

To address the aforementioned challenges, we employed a carefully designed pairwise ranking loss to learn the ordinal relations between the NREM sleep stages and between the wake and REM stages, enabling flexible sleep depth annotation within the same stage. This approach will yield a continuous value from 0 to 1, with the larger value indicating deeper sleep for each 30-s PSG epoch. Extensive validations showed that the decrease in sleep depth index is closely associated with the increase in the duration of

arousal in the next 30 seconds (Pearson correlation coefficient >0.99). Taking the unique role of REM in sleep medicine<sup>37–39</sup> into consideration, we coupled the sleep depth annotation with a REM classification for comprehensive sleep profiling. Then we inspected the nuanced differences in sleep depth index across the same and varied sleep stages, presenting that the same sleep stage could be better distinguished by sleep depth index instead of sleep staging. The whole-night continuous sleep depth index, as a type of time series, allows the extraction of various features as novel sleep biomarkers. We used the Gaussian mixture model to obtain two clusters with the extracted digital biomarkers, namely the normal sleep subtype and the disturbed sleep subtype. Subsequent analyses showcased that the disturbed sleep subtype was associated with several poor health conditions including hypertension, sleep apnea, and so on. Figure 1 displays an overview of the study.

In summary, our contributions are as follows:

- We propose a first-of-its-kind deep learning method to annotate the sleep depth index using the PSG data and existing sleep staging labels in an end-to-end way. The model structure supports scalable training on large-scale sleep data. We also deploy an easy-to-use web application for automatic annotation of the sleep depth index.



**Fig. 1 | General overview of the study.** a Four channels of physiological signals from PSG, EEG, EOG, EMG, and ECG were used in this study. The MESA, MROS, and CFS cohorts were used as the training set and interval validation set and the SHHS cohort was used as the external validation set. b Using a deep learning method, the neural network was able to transform the discrete sleep staging into a continuous sleep depth index. c There were several interesting pieces of evidence found in the sleep depth index, which added new insights into the understanding of sleep structure. Digital biomarkers extracted from the sleep depth index were used for

clustering, resulting in sleep subtypes exhibiting varied health outcomes. A web app was provided for continuous SDI annotation supporting EDF-format inputs. PSG Polysomnography, EEG Electroencephalography, EOG Electrooculography, EMG Electromyography, ECG Electrocardiography, SDI Sleep depth index, RB Ratio below a certain threshold, CV Coefficient of variation, AP Proportion of area under the sleep depth index curve, SK Skewness, MDR Mean depth value of the REM sleep epochs, PR Proportion of REM to the total sleep duration, APPE Approximate entropy, DETRf Detrended fluctuation analysis.

- Experiments on large-scale sleep cohorts and external validations demonstrated the effectiveness of our method. The decrease in the sleep depth index was strongly correlated to the increase in the duration of arousal, and the sleep depth index presented more nuanced sleep structures than conventional sleep staging.
- The whole-night sleep depth index time series yielded novel digital sleep biomarkers that were used for clustering. The resultant disturbed sleep subtype was significantly associated with a higher prevalence of several poor health conditions and an increased risk of all-cause mortality and fatal coronary heart disease.

## Results

### Data curation and deep learning model development

In this study, we mainly aimed to use a deep learning method to annotate continuous sleep depth using the existing sleep staging labels. Specifically, EEG, Electromyography (EMG), Electrooculography (EOG), and Electrocardiography (ECG) were extracted from the PSG as the input physiological signals for the model. We then converted sleep staging results labeled by the clinicians to ranks to compute the ranking loss and utilized the REM label for the REM classification loss. The MESA, MROS, and CFS cohorts were used as the training set of the deep learning model, comprising 3984 participants, and the other 1708 participants were organized into the internal validation set. The data from the SHHS cohorts was not included in the model training and thus acted as the external validation set. When trained, the deep learning model enables the transition from the discrete sleep stages to the continuous sleep depth index ranging from 0 to 1. We then analyzed the results related to the learned sleep depth index in the following sections.

### Distribution of sleep depth index and concordance across sleep stages

Supplementary Fig. 1 presents boxplots showing the distribution of the sleep depth index across the five sleep stages for each cohort. For the W stage, most sleep depth indices were below 0.1. Conversely, the deepest stage, N3, generally exhibited the highest sleep depth values, consistent with conventional expectations. Interestingly, some high sleep depth values also appeared in the N2 stage and the REM stage. The N1 stage, representing a deeper sleep than wakefulness, showcased sleep depth values mostly ranging between 0.1 and 0.5. Both the N2 and REM stages displayed a wide range of sleep depth indices, indicating varying sleep structures within the same annotated stage. In Supplementary Table 1, the first column showed the Spearman's rank correlation coefficient between the sleep depth index and the four sleep stages (W, N1, N2, and N3), which were encoded as progressively deeper sleep transitions (REM stage was not incorporated since its ordinal relation to other NREM stages could not be ascertained). The correlations, all exceeding 0.85, demonstrated good averaged concordance between the sleep depth index and traditional sleep staging results.

To comprehensively showcase whole-night sleep profiles, we have integrated sleep depth annotation with the REM classification. The classification results, detailed in Supplementary Table 1, were presented as the area under the receiver operating characteristic (AUROC) values. Overall, the micro-averaged AUROC for the four cohorts was 0.978, with a 95% confidence interval (CI) ranging from 0.977 to 0.979. For the three internal testing sets, the AUROC values were as follows: 0.990 (95% CI [0.988,0.991]) for the MESA dataset, 0.984 (95% CI [0.981,0.986]) for the MROS dataset, and 0.985 (95% CI [0.981,0.989]) for the CFS dataset. On the external validation SHHS dataset, the AUROC value was 0.975 (95% CI [0.974,0.976]), demonstrating the robustness of the REM classification part and its ability to generalize well across different datasets. Notably, the third column in Supplementary Table 1 presents the REM classification without training together with the sleep depth annotation task, where the classification performance was universally slightly worse than those in the second column, suggesting that the joint training model structure could enhance the representation learning of PSG data.

### Case studies for nuanced sleep structures shown in sleep depth index while not in traditional sleep staging

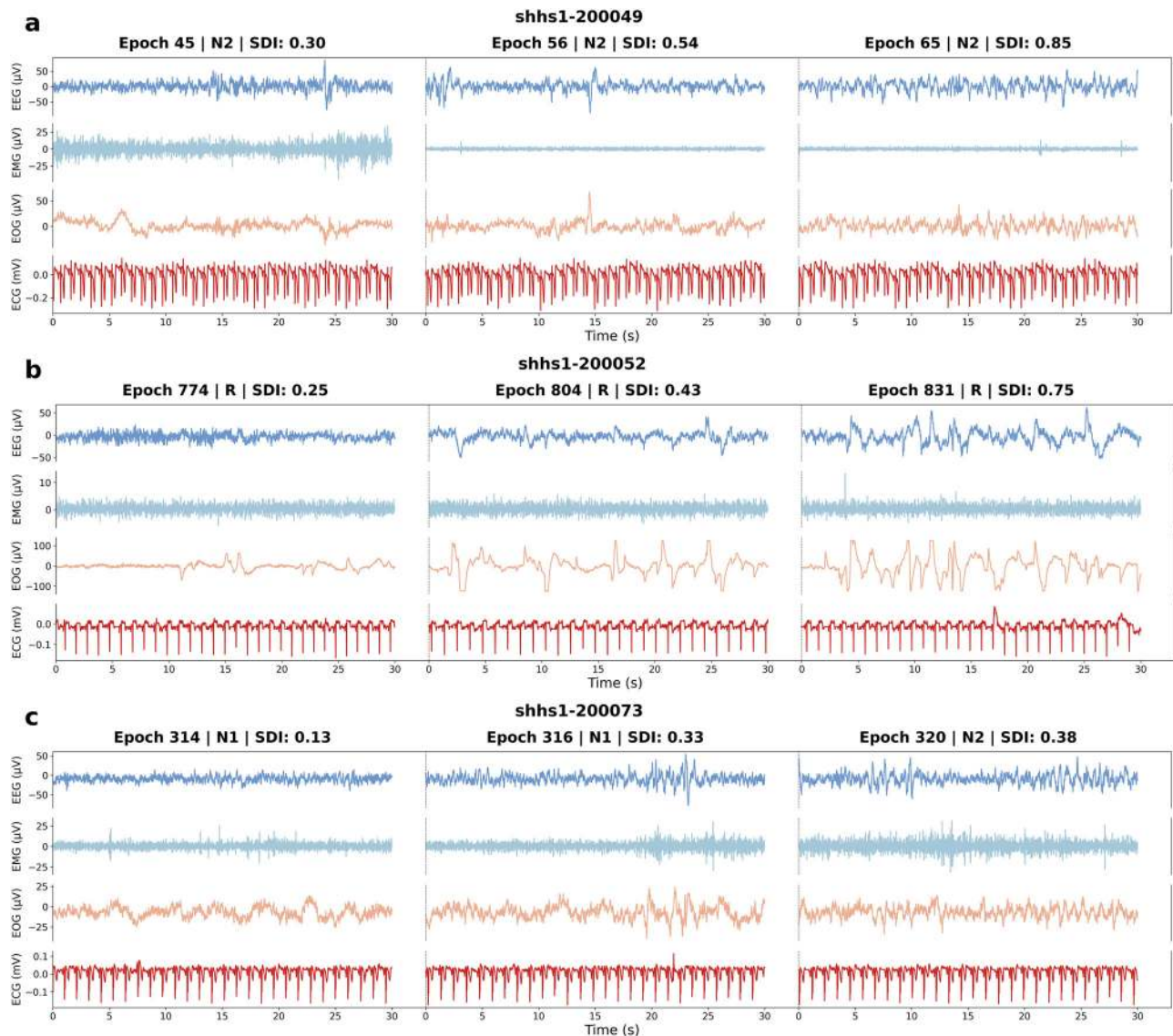
In this section, we investigated the nuanced sleep structures and patterns captured by the proposed sleep depth index, while overlooked by the traditional sleep staging methods. In Fig. 2a, the three 30-second epochs were all labeled with the N2 stage by the clinicians. However, apparently different sleep patterns could be noticed. The first epoch featured high-frequency EEG and high-amplitude EMG, resembling a waking state but also showing a representative N2 stage K-complex. The second epoch presented significantly lower EMG amplitude and more low-frequency and high-amplitude EEG waves, indicating deeper sleep states. The third epoch was labeled with a larger sleep depth index, where more low-frequency EEG patterns were observed. Furthermore, in Fig. 2b, the three 30-s epochs were all labeled as the REM stage but characterized with varied sleep depth index by our model. We were able to observe more lower-frequency and higher-amplitude EEG features in the epochs with larger sleep depth values. In addition, larger sleep depth indexes were associated with more evident eye movements showcased by the EOG. The N1 stage, which suffers from poor inter-rater reliability, has been challenging for existing sleep staging methods due to the subtle differences between N1 and N2. In Fig. 2c, although both the first epoch and the second epoch were recognized as N1, the sleep depth index would assign a higher value to the second epoch in relation to the lower-frequency and higher-amplitude EEG pattern. Nevertheless, the physiological signals in the third epoch, which was labeled with the N2 stage, resembled those from the second epoch and thus shared a close sleep depth index.

### The decrease in sleep depth index and the increase in duration of arousal was highly correlated

Arousal during sleep represents a shift from deep sleep to light sleep or from sleep to wakefulness<sup>40</sup>. A low likelihood of arousal indicates deeper sleep depth, making it less likely for the sleeper to be awakened. For the four cohorts studied, arousal events were annotated with their start and duration times. We then defined the duration of arousal for a certain 30-second epoch as the proportion of the arousal duration within that epoch. The decrease in the sleep depth index was computed by subtracting the value at time  $t$  from the value at time  $t - 1$ . We created deciles for the annotated sleep depth index, resulting in ten equal intervals. We then averaged the arousal durations within each interval, with error bars representing two-sided confidence intervals. As shown in Fig. 3, the linear regression analysis for each cohort revealed strong correlations between the decrease in sleep depth index and the increase in the duration of arousal. Specifically, the Pearson correlation coefficients were 0.9913 for SHHS, 0.9968 for CFS, 0.9955 for MESA, and 0.9977 for MROS. It could be noted that larger decreases in sleep depth correspond to broader confidence intervals, probably attributed to the fewer PSG epochs with long arousal durations. Moreover, we investigated this linear relation with more bins (100) split shown in Supplementary Fig. 2, where prominent correlations could still be found.

### Clustering based on the digital biomarkers derived from sleep depth index resulted in two subtypes with different health conditions and outcomes

As for the whole-night continuous sleep depth index for each subject, we extracted a set of time series features as the novel digital biomarkers. In the time domain, the coefficient of variation (CV) and skewness (SK) were computed. Based on the physiological nature of the sleep depth index, we then devised several intuitive features to reflect the sleep states. The first was the ratio below a certain threshold (RB), indicating the proportion of shallow sleep during the night. We set 0.2 as the threshold in our study. The second one was the proportion of area under the sleep depth index curve (AP), computed by dividing the integration value of the sleep depth index by the total sleep duration. This feature showcased the efficiency of sleep more accurately than the conventional sleep efficiency (SE) metric, which was computed as the ratio of sleep duration to the in-bed period. The approximate entropy (APPe) and the detrended fluctuation analysis results



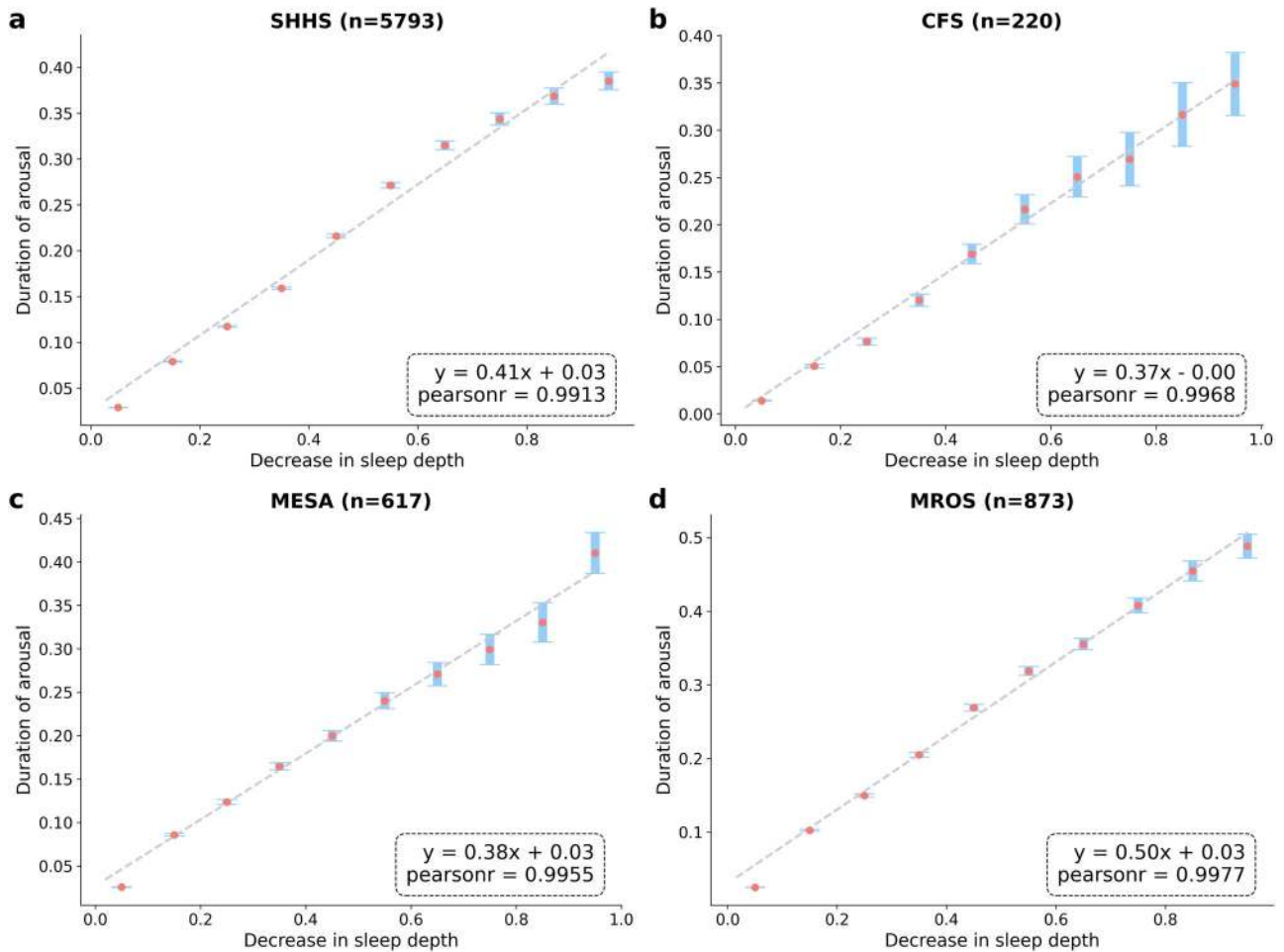
**Fig. 2 | Cases of varied patterns of the physiological signals across the sleep stages.** **a** The three stages belonged to the same N2 stage but the sleep depth values differed. The sleep depth index better captured the lower frequency feature of deep sleep and the smaller magnitude of EMG. **b** The three stages belonged to the same REM stage but the sleep depth values differed. The sleep depth index better captured the lower frequency feature of deep sleep and the larger magnitude of EOG. **c** The first two

stages belonged to the N1 stage but the second one was labeled with a slightly larger sleep depth index. The third stage shared similar patterns with the second stage, but it was labeled as the N2 stage, where the sleep depth values were close. SDI Sleep depth index, EEG Electroencephalography, EOG Electrooculography, EMG Electromyography, ECG Electrocardiography.

(DETRf) were extracted to analyze the complexity dimension of the sleep depth index. Since we have acquired the REM classification results at the same time, the mean depth value of the REM sleep epochs (MDR) and the proportion of REM to the total sleep duration (PR) were also extracted. Subsequently, these digital biomarkers were used as input for the Gaussian mixture model for clustering, resulting in two subtypes of sleep, namely the normal sleep subtype and the disturbed sleep subtype.

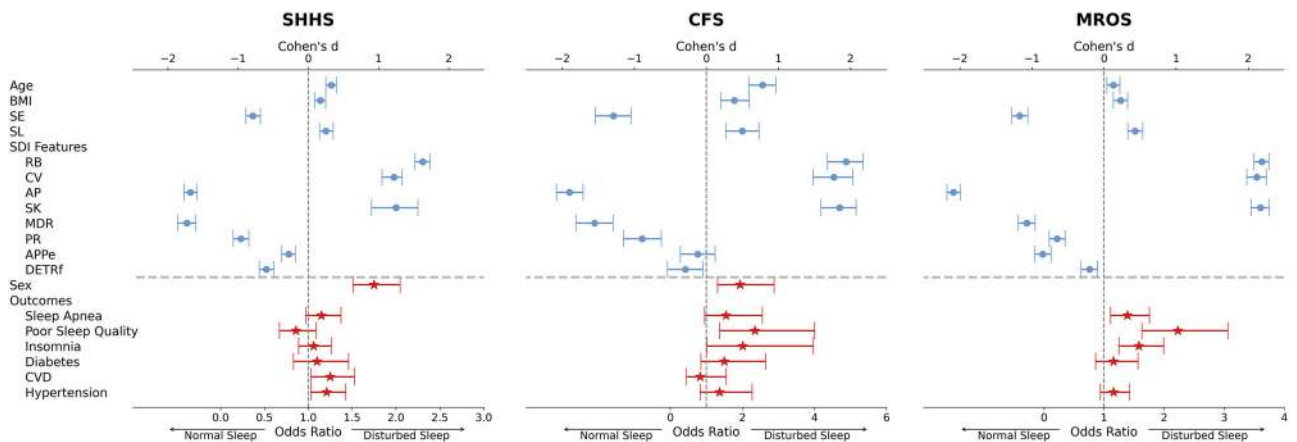
In Fig. 4, the effect sizes estimates with 95% CIs for demographic, traditional metrics, SDI-based features, and several health outcomes are displayed. The detailed values of the effect sizes and the mean values were stored in Supplementary Table 3. Generally, for the three cohorts, SHHS, CFS, and MROS, participants in the disturbed sleep group were older and had a large Body Mass Index (BMI), but this finding was only prominent in the CFS cohort. For the SHHS cohort and CFS cohort, male participants were more likely to be in the disturbed subtype. Regarding traditional sleep metrics, participants in the disturbed sleep group had significantly lower sleep efficiency, as indicated by large effect sizes. They also tended to have

longer sleep latency (SL), though this tendency was much weaker. As for the SDI features, the disturbed sleep featured larger RB and smaller AP, indicating a more dominant ratio of shallow sleep and lower sleep efficiency than the normal sleep group. The larger CV in the disturbed sleep group showcased more dispersed patterns than the normal and the larger skewness indicated more frequent or extremely high values. The normal sleep group was characterized by larger MDR and PR, presenting deeper sleep in the REM and longer duration of the REM stage. Notably, the normal sleep group presented larger APpe and DETRf, showcasing a higher level of complexity when compared with the disturbed group. There was an apparent opposite trend for the CV feature and the complexity-based features. It was suspected that the CV reflects more on the variation of regular disturbance across the night and that those with disturbed sleep patterns were frequently affected by these interruptions. On the other side, as measures of complexity, APpe and DETRf presented more nuanced sleep structures similar to the heart rate variability (HRV) for ECG data, where reduced HRV has been shown to be associated with some poor health outcomes<sup>41,42</sup>.



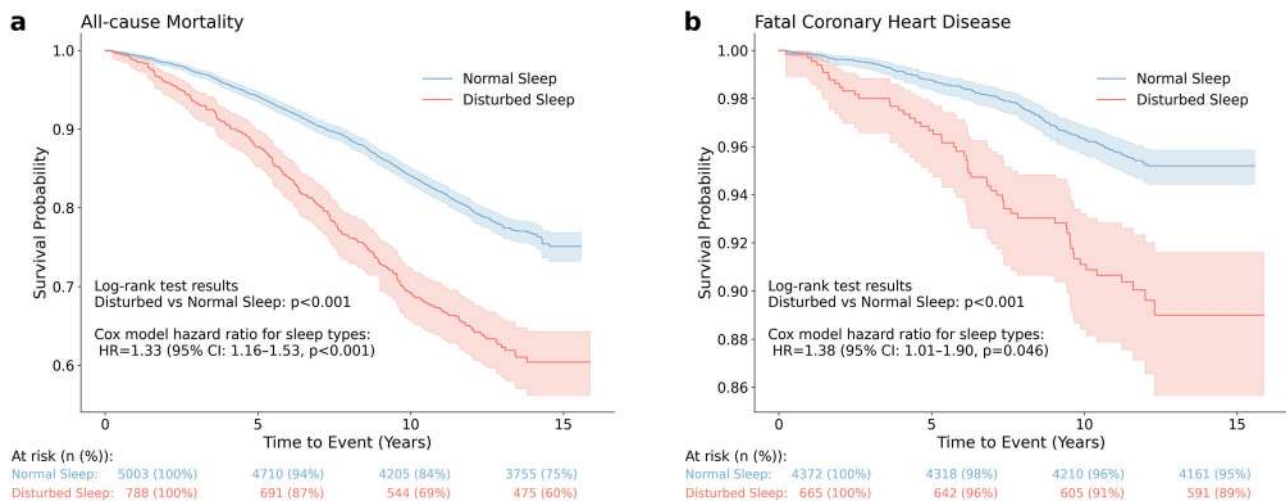
**Fig. 3 | The correlation between the decreased magnitude of the sleep depth index and the increase in the duration of arousal.** **a** The SHHS cohort. **b** The CFS cohort. **c** The MESA cohort. **d** The MROS cohort. The duration of arousal was computed as the proportion of arousal duration in a 30-s epoch. We calculated the ten deciles of 0

to 1 and averaged the values in each interval for linear regression fitting. The red dotted line represented the diagonal line. The average relationship was almost perfectly linear.



**Fig. 4 | Effect size estimates with 95% CIs for demographic, SDI-based features, and several health outcomes.** The region above the dashed line with blue data points was about the comparison of continuous variables, using the t-test to compare the between-group differences and Cohen's d as the measure of effect size. The effect size was computed by subtracting the value of the normal sleep group from the disturbed sleep group. The lower region was about the comparison of categorical variables. For sex, males were coded as 1 and females as 0. For outcomes, the presence of a specific outcome was recorded as 1. The Chi-squared test was used to compare the between-group differences for the sex variable and odds ratio was utilized as the measure of effect

size. Logistic regression was used to estimate the odds ratio of the outcome variables, adjusting for age, BMI, sex, SE, and SL. The disturbed sleep group was regarded as value 1 when computing the effect size. BMI Body Mass Index, SE Sleep efficiency, SL Sleep latency, SDI Sleep depth index, RB Ratio below a certain threshold, CV Coefficient of variation, AP Proportion of area under the sleep depth index curve, SK Skewness, MDR Mean depth value of the REM sleep epochs, PR Proportion of REM to the total sleep duration, APPE Approximate entropy, DETRf Detrended fluctuation analysis, CVD Cardiovascular disease.



**Fig. 5 | Survival outcomes by clinical subtype.** Kaplan–Meier curves across the two subtypes for **a** all-cause mortality **b** fatal coronary heart disease. HR hazard ratio, CI confidence interval.

For the two subtypes, the differences in the prevalence of several health were investigated in the lower region of Fig. 4 labeled with red. Logistic regression controlling for age, BMI, sex, SE, and SL was used to estimate the odds ratio (OR) with 95%CI. Sleep apnea was defined as having an Apnea-Hypopnea Index (AHI) larger than 5. The results of poor subjective sleep quality and insomnia were sourced from the morning surveys of the corresponding cohorts. Note that there was no outcome of cardiovascular disease (CVD) for the MROS cohort. We could see from Fig. 4 that the disturbed group across the three cohorts showcased a significantly higher prevalence of sleep apnea, with the odds ratio ranging from 1.15 to 1.55. Poor subjective sleep quality was also significantly associated with the disturbed subtype, with the odds ratio being 2.35 (95% CI 1.37–4.0) for the CFS cohort, and 2.23 (95% CI 1.63–3.06) for the MROS cohort, but not significant for the SHHS cohort. Insomnia occurred more frequently for participants in the CFS and MROS cohort with the disturbed subtypes, where the odds ratios were 2.01 and 1.58. The difference in the prevalence of diabetes between the two subtypes was not statistically significant across the three cohorts. CVD was more prevalent in the disturbed subtype than the normal subtype for the SHHS cohort (OR 1.25 95% 1.03–1.53), but not statistically significant for the CFS cohort. Populations belonging to the disturbed subtype were associated with a higher prevalence of hypertension for the SHHS (OR 1.21 95% 1.03–1.43).

As for the SHHS cohort, time-to-event data was extracted for survival analysis of the clustered two subtypes. In Fig. 5, Kaplan–Meier survival estimates provided a visual interpretation of the crude probability of all-cause mortality and fatal coronary heart disease. Log-rank tests showcased significant differences between the two subtypes in the survival probability ( $p < 0.001$  for both all-cause mortality and fatal coronary heart disease). From the results, participants in the disturbed subtype had a significantly lower survival probability than the normal subtype. The Cox regression model was then used to determine the hazard ratios (HRs) and 95% confidence intervals to compare the risks in the normal sleep and disturbed sleep groups, adjusting for age, BMI, sex, SE, and SL. For all-cause mortality, the disturbed sleep group was associated with a 33% (hazard ratio 1.33, 95% CI 1.16–1.53,  $p < 0.001$ ) increased risk compared with the normal sleep group. For the fatal coronary heart disease, participants with the disturbed subtype suffered a 38% (hazard ratio 1.38, 95% CI 1.01–1.90,  $p = 0.046$ ) increased risk with respect to the other normal subtype.

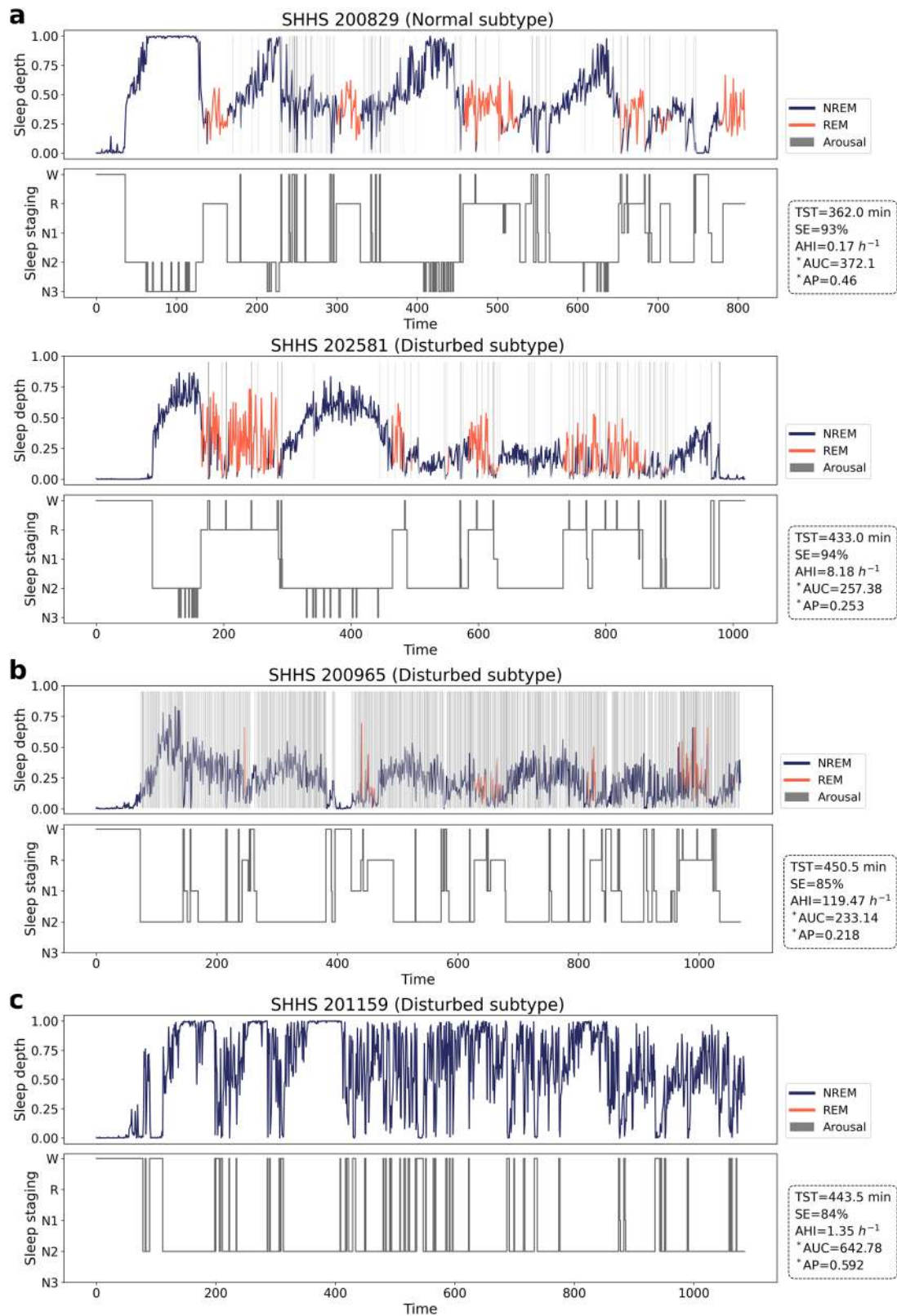
In Fig. 6, we investigated the differences in SDI patterns for the two clustered subtypes with some instances. We presented the full-night SDI annotations and hypnograms for four SHHS patients. The boxes on the right displayed some metrics such as total sleep time (TST), sleep efficiency, AHI, area under the sleep depth index curve (AUC), and AP. Specifically, for

Fig. 6a, two participants with similar structure hypnograms but belonging to varied clustered sleep subtypes were compared. The participant with disturbed sleep was labeled with a sleep efficiency of 94%, a little larger than the one displayed in the upper sub-figure of 93%. However, the latter showcased a larger AP (0.46) than the former (0.253), and in the plots of SDI, we could observe that the participant with disturbed sleep had very shallow sleep in the second half of the night, which would not be discovered by the traditional sleep staging method. In Fig. 6b, a patient characterized by constant arousal and severe sleep apnea featured shallow sleep across the whole night, ending with a very low AP (0.218) and AUC (233.14) but still a high sleep efficiency (85%) and high total sleep time (450.5 min). Next for the participant with the disturbed subtype as shown in Fig. 6c, no REM or N3 were identified, and no sleep cycles were observed. Though the participants had a high AUC and AP, the sleep pattern fiercely fluctuated from deep sleep to shallow and vice versa, which was less obvious in the frequent transitions between the wake and N2 stages as shown in the hypnogram.

## Discussion

In this study, we developed and externally validated a deep-learning method for end-to-end annotation of the sleep depth index using a large scale of 11485 PSG recordings. We first viewed the basic properties of the sleep depth index in the second part of the results, the distribution of SDI and its relations to sleep staging. Then in the next part, we checked the correlation between the SDI and the arousal event, which is an indispensable element when researching the sleep fragmentation problem. In the fourth part, we investigated the specific cases showing that SDI would be better than sleep staging in capturing certain nuanced sleep patterns. Finally, in the last section of results, a range of comparisons were made to validate that digital biomarkers extracted from the whole-night SDI had the potential to promote clustering and that different subtypes were associated with significantly varied health conditions and outcomes.

There were some trends of SDI that showcased intuitive findings with respect to the sleep stages. For instance, SDIs in the W and N1 stages were generally small and those in the N3 stages were significantly large. The difference in SDI for epochs in the N2 stage varied drastically, and sometimes the boundary between N1/N2 and N2/N3 was not apparent but the SDIs were close. Unlike traditional sleep staging, which assigns five coarse-grained classes to the sleep state, the proposed sleep depth index offers a more nuanced description of sleep structure, capturing varied patterns of physiological waveform that are not discernible through conventional sleep staging, as shown in Fig. 2. This variability of sleep depth index is meaningful as it reflects the intricate transitions in sleep states, as demonstrated in our case studies. Given the distinctiveness of the REM stage<sup>37–39</sup>, we integrated



**Fig. 6 | Plots of the whole-night SDI along with the expert-labeled hypnogram, where the degree of the gray color indicates the proportion of arousal in a 30-s epoch. In the information box on the right, sleep metrics with \* are derived from the sleep depth index. a** Two participants belonging to two different sleep subtypes had close sleep efficiency metrics but significantly different AUC and AP. **b** A patient with severe sleep apnea featuring constant shallow sleep and sleep fragmentation.

**c** The participant with disturbed sleep, though presented a relatively high AP, suffered from excessive sleep depth fluctuation. REM Rapid eye movement, NREM, Non-rapid eye movement, TST Total sleep time, SE Sleep efficiency, AHI Apnea Hypopnea Index, AUC, Area under the sleep depth index curve, AP Proportion of area under the sleep depth index curve.

sleep depth annotation with REM stage classification, producing comprehensive sleep profiling. The predicted REM would also serve as a supplement to feature engineering in addition to the SDI. The proposed SDI was not purposed to replace the sleep stages completely but acted as a valuable source for sleep clinicians to use.

In sleep medicine, arousal indicates a transition from deep to light sleep, making it a suitable candidate for measuring sleep depth. Our experiments showed that a decrease in the sleep depth index strongly correlated with an increase in the duration of arousal, suggesting its potential for monitoring sleep disturbance. More importantly, the arousal event was routinely labeled by experienced clinicians in a cumbersome manner and might suffer from inter-rater variability. As an automatic approach, the proposed method presents a more consistent way than human labeling and adds new insights to measure sleep fragmentation.

Extracting the whole-night SDI for an individual results in a time series rich with information about sleep conditions. Time-domain and Nonlinear features could be directly computed but we could mine more intuitive ones such as AP, which is capable of introducing more nuanced findings than the classic sleep efficiency metric. Based on these digital biomarkers derived from SDI, we performed clustering to obtain two sleep subtypes featuring significant patterns in sleep health. Participants in the identified disturbed subtype were more associated with bad health conditions such as CVD, hypertension, sleep apnea, and insomnia. The disturbed sleep subtype had a higher prevalence of all-cause mortality and fatal coronary heart disease, which offered meaningful indications for clinical practice and relevant prevention. Notably, these are not all the digital biomarkers we can extract. We then envision that more subsequent studies could be conducted to investigate the effects of various time-series features that are able to be explored by the sleep depth index, thus yielding more novel digital biomarkers for sleep medicine. These digital biomarkers would not totally make the traditional sleep metrics such as TST and SE displaced but would help to explore more sleep patterns that were previously overlooked.

Although the concept of sleep depth is frequently discussed in both public and clinical contexts, there are no widely accepted quantitative measurements for sleep depth. Traditional sleep medicine primarily considers the N3 stage as deep sleep<sup>43</sup> but a single discrete class is far from enough to precisely reflect a high degree of sleep depth. Recently, the ORP<sup>23</sup> index had demonstrated utility in clinical applications<sup>24–28</sup>, but it relied solely on EEG and required manual computations by experts, limiting its efficiency and accessibility. Moreover, the method was not open-sourced mainly due to the unavailability of the reference table to rank the EEG power values, hindering the research community from conducting related research. Consequently, based on the abundant off-the-shelf sleep staging labels, we proposed to use deep learning, with its extraordinary performance in clinical medicine<sup>44–49</sup>, for annotating the sleep depth index on large-scale PSG in an end-to-end way. Nevertheless, the native sleep staging labels are coarse-grained and discrete in five classes, prohibiting direct supervised training. To tackle this, we designed a pairwise ranking loss to effectively learn the ordinal relations between stages and take the uncertainty of the relation between the REM stage and other NREM stages into account, making it the first attempt to produce a continuous sleep depth measure from sleep staging labels. The model was based on the Transformer structure<sup>50</sup>, which is most famous for its scalability of training large-scale neural networks that might show emerging abilities<sup>51–53</sup>. We envision that our proposed method of annotating SDI would add new insights into clinicians' interpretations of the PSG besides the traditional sleep analyses. It could be an important quantitative measurement of sleep depth as it is automatically output by the machine learning model instead of human labeling which might suffer from inter-rater inconsistency.

An easy-to-use web application for automatic annotation of the sleep depth index was deployed online for the evaluation of the polysomnography data. The demonstration is shown in Supplementary Fig. 3. Users can upload EDF files with labeled channel names to obtain analysis results and visualizations. We are continuously enhancing this application to offer more features.

A limitation of our method is its reliance on four PSG channels (EEG, EMG, EOG, and ECG). Future studies should investigate whether similar performance can be achieved with fewer channels, making the approach more feasible in remote and underserved areas where full PSG monitoring is not available. Previous research has shown the effectiveness of using fewer physiological signals in wearable devices for sleep staging<sup>54,55</sup>, suggesting the potential for portable sleep depth annotation systems. Although we have mentioned that a better sleep state can be explored in full PSG monitoring, wearable devices may be a good trade-off in certain medical scenarios. The other issue with the proposed method is that we only include a single channel of EEG, EMG, EOG, and ECG in modeling, thus a question comes out as to whether the performance would be further improved by incorporating more channels or taking the respiratory signals into consideration. Moreover, the model could be tested on a larger dataset with an increasing number of parameters to further verify its scalability. Our current model used a 30-s scale for annotation, based on available sleep staging labels. Higher resolutions could be achieved by modifying model structures and loss functions<sup>56,57</sup>, which is a direction for future research.

In summary, our study underscored the potential of the proposed sleep depth index annotation method as a valuable ancillary tool in clinical sleep medicine. The resultant sleep depth index promises several utilities in real clinic practice and can be explored to yield novel digital biomarkers for sleep health. We hope this work will inspire further research into AI's role in sleep depth annotation and its significance in sleep medicine.

## Methods

### Dataset and preprocessing

Four large-scale cohorts were used in this study. From the NSRR website<sup>58</sup>, the Sleep Heart Health Study (SHHS) is a multi-center cohort study implemented by the National Heart Lung & Blood Institute to determine the cardiovascular and other consequences of sleep-disordered breathing<sup>59</sup>. The Cleveland Family Study (CFS) is the largest family-based study of sleep apnea worldwide, which was begun in 1990 with the initial aims of quantifying the familial aggregation of sleep apnea<sup>60</sup>. Multi-Ethnic Study of Atherosclerosis (MESA) is an NHLBI-sponsored 6-center collaborative longitudinal investigation of factors associated with the development of subclinical cardiovascular disease and the progression of subclinical to clinical cardiovascular disease<sup>61</sup>. MrOS is an ancillary study of the parent Osteoporotic Fractures in Men Study<sup>62</sup>. For the SHHS and MROSA cohorts, we used records belonging to visit-1. The basic statistics are shown in Supplementary Table 2. We used the MNE library<sup>63</sup> to extract the PSG signals from the raw files and resample them to 100 Hz for concordance. The signals were then trunked to 30-second epochs and saved with the corresponding sleep staging labels and arousal annotation. We chose one EEG (C4) channel, one EMG (chin), one EOR (right eye), and one channel of ECG. We split training sets and internal testing sets for the CFS, MESA, and MROSA cohorts by a 7:3 ratio. The SHHS cohort was left as the external validation dataset.

### Definitions of sleep apnea, sleep quality, insomnia, and mortality events

For the four datasets, Apnea-hypopnea index (AHI) values, computed as (All apneas + hypopneas with  $\geq 30\%$  nasal cannula [or alternative sensor] reduction with  $\geq 4\%$  oxygen desaturation) / hour of sleep, were extracted from the data harmonized by the NSRR team. Then sleep apnea was defined as AHI  $\geq 5$ . At the time of these PSG studies, subjects were required to complete some morning surveys with respect to their last night's sleep. Note that there are some surveys inquiring about the sleep habits of the participants but we did not use them since we concentrated on the immediate sleep feeling after the PSG studies. In our experiment, three variables *diffa10*, *ltdp10*, *rest10* were selected for the SHHS dataset. Specifically, the *diffa10* variable measured the difficulty of falling asleep, with 0 indicating No and 1 indicating Yes. Thus participants with *diffa10* being 0 were defined as having insomnia disorders. The *ltdp10* variable measured the quality of sleep in terms of light or deep sleep, with five degrees where a value of 1 indicated

light sleep and a value of 5 indicated deep sleep. The *rest10* variable measured the quality of sleep in terms of restless or restful, with five degrees where a value of 1 indicated restless sleep and a value of 5 indicated restful sleep. We selected degree 1/2 and degree 4/5 of each index to compare the results. Participants with both the *ltdp10* and *rest10* being less than 2 were defined as having poor sleep quality and those with both the *ltdp10* and *rest10* being larger than 4 were defined as having good sleep quality. As for the CFS dataset, we extracted four variables *easlp*, *difbak*, *slpqua*, *desslp*. The *easlp* variable measured how easy is it for the subject to fall asleep last night, which was rated on a scale of 1–6, with 1 being very easy and 6 being not at all easy. The *difbak* variable indicated the difficulty of falling back to sleep, with value 0 showing No and value 1 showing Yes. CFS participants with *easlp* less equal to 3 and *difbak* being 0 were defined as not having insomnia and those with *easlp* more than 4 and *difbak* being 1 were defined as having the insomnia symptom. The *slpqua* variable measured the subjects' feeling of the overall quality of sleep last night, which was rated on a scale of 1–6, with 1 being extremely refreshing and 6 being not refreshing. The *desslp* variable was about the participants' description of sleep last night, with the value 1 being excellent, 2 being very good, 3 being fair and 4 being poor. Then, participants with *slpqua* being less than 3 and *desslp* less than 2 were defined as having good sleep quality, and those with *slpqua* being more than 4 and *desslp* being more than 3 were defined as having poor sleep quality. As for the MROS dataset, three variables *poxfall*, *poxqual1*, *poxqual3* were selected. The *poxfall* variable was a numeric index measuring how long it took for the participant to fall asleep at bedtime last night. MROS participants with *poxfall* more than 40 min were defined as having the insomnia symptom. The *poxqual1* variable inquired about the participants' feelings about the sleep being light or deep, with five degrees similar to the *ltdp10* variable of the SHHS dataset. The *poxqual3* measured the feelings about the sleep being restless or restful, with five degrees similar to *rest10* in SHHS. MROS participants with both *poxqual1* and *poxqual3* being less than 2 were defined as having poor sleep quality and those with both *poxqual1* and *poxqual3* being larger than 4 were defined as having good sleep quality. Participant deaths in the SHHS cohort, which were identified and confirmed using multiple concurrent approaches including follow-up interviews, written annual questionnaires, telephone contacts, and so on<sup>64</sup>, were sourced from the NSRR dataset. The fatal coronary heart disease was recorded in parent studies datasets at the NSRR website. The censoring time (days to most recent contact or death) was also extracted accordingly.

### Definitions of digital sleep biomarkers

We extracted digital biomarkers from the whole-night sleep depth index from the time domain and the non-linear perspective. As for the time domain, the coefficient of variation and skewness were investigated. Specifically for SDI-featured ones, a ratio below 0.2 (RB, the ratio of sleep depth index less than 0.2), and the proportion of area to the total sleep time (AP) were extracted. As for non-linear features, approximate entropy and detrended fluctuation analysis were used. Approximate entropy is a technique used to quantify the amount of regularity and the unpredictability of fluctuations over time-series data. Smaller values indicate that the data is more regular and predictable<sup>65</sup>. The detrended fluctuation analysis is a method for determining the statistical self-affinity of a signal, which has been widely used in the physiological domain<sup>66</sup>.

### Statistical analysis

Two-sided t-tests were conducted on the between-group differences of the two clustered subtypes in demographic features, traditional metrics, and digital biomarkers from the SDI. Cohen's d was used as a measurement of the effect size. Chi-squared tests were utilized to compare the between-group differences for the sex feature, with the odds ratio used to measure the effect size. For the health outcomes, the odds ratio was computed by a Logistic model adjusting for age, sex, BMI, SE, and SL. Bootstrapping was implemented to compute the 95% confidence intervals for the estimates. The log-rank test was used to compare the between-group differences in the survival probability for the two subtypes. The Cox regression model

was used to compute the hazard ratios, adjusted for age, sex, BMI, SE, and SL.

### Model structure

An overview of the model is depicted in Supplementary Fig. 4. First, the raw input PSG was segmented into a sequence of patches. Specifically, we had an input PSG epoch  $x \in \mathbb{R}^{C \times L}$ , where  $C$  was the number of physiological channels and  $L$  was the sequence length. In our study, data from four channels were collected in 30-s epochs at a sampling frequency of 100 Hz for predicting sleep states, thus  $C$  was 4 and  $L$  was 3000. The sequence of each channel was first split into  $N_c$  fixed-size patches with patch size  $P$  ( $N_c = L/P$ ). Then the patches from different channels were flattened, yielding a 1D vector composed of  $N = C \times N_c$  patches. The flattened patch vector was projected to  $D$  dimensions through a trainable linear projection. The outputs of this projection were conventionally referred to as patch embeddings. Following ideas of the original Vision Transformer (ViT) architecture<sup>52</sup>, learnable and randomly initialized positional embeddings were added to the projected patch embeddings to provide the model with information about the position of the patches in the PSG. In addition, we added the channel embeddings since for the PSG every channel has the signal modality varying dramatically and recent works have shown the necessity to take this into consideration<sup>53,67</sup>. We then prepended a learnable *CLS* token to the patch embeddings to represent the global contextual information learned by the model. The final embedding vectors then served as input of the standard Transformer encoder which consisted of alternating layers of multihead self-attention (MSA) and (multilayer perceptron) MLP blocks<sup>50</sup>, where LayerNorm (LN)<sup>68</sup> was applied before every block, and residual connections<sup>69</sup> after every block. The standard self-attention mechanism allowed the model to weigh the importance of different patches relative to each other. With  $Q, K, V$  being the query, key, and value matrices linearly projected from the input embedding vector  $X \in \mathbb{R}^{N \times D}$ ,

$$[Q, K, V] = XU_{qkv}, \quad U_{qkv} \in \mathbb{R}^{D \times 3D_h} \quad (1)$$

the attention scores were computed as follows:

$$A = \text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{D_h}} V \right) \quad (2)$$

where  $A_{ij}$  was computed based on the respective  $Q_i$  and  $K_j$  representations and  $D_h$  was computed as  $D/m$  with  $m$  being the number of heads for multihead self-attention. Specifically as for the multihead self-attention, it was an extension of the native self-attention in which  $m$  self-attention operations (heads) are conducted. Then these outputs were concatenated and projected to the  $D$  dimension

$$\text{MSA}(X) = [A_1(X); A_2(X); \dots; A_k(X)]U_{msa}, \quad U_{msa} \in \mathbb{R}^{m \cdot D_h \times D} \quad (3)$$

The MLP block consisted of two linear layers with a GELU non-linearity. So the  $l$ th encoder function could be written as

$$X'_l = \text{MSA}(\text{Layer Norm}(X_{l-1})) + X_{l-1}, \quad l = 1 \dots L_T \quad (4)$$

$$\text{Encoder}(X_l) = \text{MLP}(\text{Layer Norm}(X'_l)) + X'_l, \quad l = 1 \dots L_T \quad (5)$$

where  $L_T$  was the number of layers of the Transformer encoder. In this study, patch size was set to 100, projection dimension  $D$  to 512, encoder depth to 6, heads number to 8, and MLP dimension to 2048. After passing through the transformer block, the encoded *CLS* embedding was separately fed into a MLP for predicting the sleep depth and another MLP for predicting the REM stage, both of which resembled the Transformer's MLP block. The outputs of the sleep depth MLP were scalars with continuous values annotating the sleep depth, and the outputs from the REM MLP were vectors in length of 2 classifying whether the input epoch represented a REM

stage. Note that during training the outputs of the depth head were not bounded and we post-processed them with a Sigmoid function to make them have values ranging between 0 and 1. The reason why we did not bound the range in the depth head was that we observed better results obtained in this experiment setting. The pair rank loss and cross-entropy loss were computed on the sleep depth annotation and REM binary classification respectively with details in the following section.

As for the hyperparameter choices, the patch size of 100 was chosen based on the fact that the majority of wave patterns (e.g., K-complex, sleep spindles) lasts about 0.5–2s and the sampling rate is 100 Hz so that the model is capable of capturing enough features. The projection dimension, encoder depth, number of heads, and other hyperparameters were determined mainly in light of previous relevant studies. In our preliminary experiments, we also found no significant changes in the model’s performance in reflecting the clinical indication.

**Loss function**

For the design of the loss function for sleep depth annotation, we aimed to leverage the ordinal relationships between different sleep stages, while accounting for uncertain relationships between specific sleep stages. Given a batch of predicted sleep depths  $\mathbf{p} = [p_1, p_2, \dots, p_n]$  and corresponding true labels  $\mathbf{y} = [y_1, y_2, \dots, y_n]$ , the goal was to minimize the loss that considered the margin between different pairs of sleep stages and penalizing the for incorrect ordinal relations. Let  $\mathcal{M}$  be the mapping from pair types to margins. For a pair of sleep stages  $(i, j)$ , the margin was denoted by  $\mathcal{M}_{ij}$  and computed as:

$$\mathcal{M}_{ij} = \begin{cases} 1 & \text{if } (i, j) = (0, 1) \text{ or } (i, j) = (1, 0), \\ 0.5 & \text{if } (i, j) = (1, 2) \text{ or } (i, j) = (2, 1), \\ 1.5 & \text{if } (i, j) = (2, 3) \text{ or } (i, j) = (3, 2), \\ 1.2 & \text{if } (i, j) = (0, 4) \text{ or } (i, j) = (4, 0), \end{cases} \quad (6)$$

$\mathbf{y} \in \{0, 1, 2, 3, 4\}^n$  was the true sleep stage label for the batch of samples, where 0/1/2/3/4 corresponded to W/N1/N2/N3/R. The set of uncertain relationships was denoted by  $\mathcal{U}$ . First, we needed to compute all possible pairs from the predicted depths and true labels, where  $\mathbf{p}_{\text{pairs}} = \{(p_i, p_j) | i \neq j\}$ ,  $\mathbf{y}_{\text{pairs}} = \{(y_i, y_j) | i \neq j\}$  with each pair represented as  $(p_i, p_j)$  and  $(y_i, y_j)$ . Second, for each pair of true labels, we retrieved the corresponding margins,  $\mathcal{V}_{ij} = \mathcal{M}(y_i, y_j)$  for  $(y_i, y_j) \in \mathbf{y}_{\text{pairs}}$ . In this study, uncertain relationships were identified as  $\mathcal{U} = \{(1, 4), (4, 1), (2, 4), (4, 2), (3, 4), (4, 3)\}$  since we were uncertain about the relation between the REM stage and the other three NREM stages. The mask was formulated as an indicator function as  $\mathbb{I}[(y_i, y_j) \in \mathcal{U} \text{ for } (y_i, y_j) \in \mathbf{y}_{\text{pairs}}]$ .

Penalties were computed based on the predicted depths and margins:

$$\mathcal{P}_{ij} = \max(0, V_{y_i, y_j} - \text{sign}(y_i - y_j)(p_i - p_j)) \quad (7)$$

where sign was the sign function. We then computed the final pair rank loss by averaging the penalties, ignoring the uncertain relationships:

$$\mathcal{L}_{rank} = \frac{1}{\mathcal{N}} \sum_{(i,j) \in \mathbf{y}_{\text{pairs}}} \mathcal{P}_{ij} \cdot (1 - \mathbb{I}[(y_i, y_j) \in \mathcal{U}]) \quad (8)$$

, where  $\mathcal{N} = \sum_{(i,j) \in \mathbf{y}_{\text{pairs}}} (1 - \mathbb{I}[(y_i, y_j) \in \mathcal{U}])$

On the other side, we used the cross-entropy loss for the REM prediction task. To do this, we need to convert the five-class sleep staging labels to binary labels, we defined  $y_i^* = \mathbb{I}(y_i = 4)$ , then the cross-entropy loss was:

$$\mathcal{L}_{CE} = - \sum_i [y_i^* \log(\hat{y}_i) + (1 - y_i^*) \log(1 - \hat{y}_i)] \quad (9)$$

where  $y_i$  was the true label and  $\hat{y}_i$  was the corresponding predicted probability from the REM classification module. The final loss function for the

joint model was formulated as:

$$\mathcal{L} = \mathcal{L}_{rank} + \alpha \mathcal{L}_{CE} \quad (10)$$

where  $\alpha$  was a hyperparameter of loss weight. The overall training objective was to minimize the combined loss  $\mathcal{L}$ . In this study, the loss weight  $\alpha$  was set to 1. The predicted sleep depth index would then be input to a Sigmoid function to make the range between 0 and 1.

**Evaluation metrics**

The Pearson correlation coefficient was used to measure the correlation between the decrease in sleep depth and the increase in the duration of arousal. The Spearman’s rank correlation was used to measure the correlation between the sleep depth and the W/N1/N2/N3 sleep stages. The area under the receiver operating characteristic (AUROC) was used to assess the classification performance for REM stage classification. Confidence intervals were computed by Bootstrapping.

**Data availability**

The datasets used in this study are available at <https://sleepdata.org/>.

**Code availability**

The source code is available at <https://github.com/sczzz3/SDI>, or <https://github.com/PKUDigitalHealth/SDI>.

Received: 2 August 2024; Accepted: 30 March 2025;

Published online: 11 April 2025

**References**

1. Tregear, S., Reston, J., Schoelles, K. & Phillips, B. Obstructive sleep apnea and risk of motor vehicle crash: systematic review and meta-analysis. *J. Clin. sleep. Med.* **5**, 573–581 (2009).
2. Smolensky, M. H., Di Milia, L., Ohayon, M. M. & Philip, P. Sleep disorders, medical conditions, and road accident risk. *Accid. Anal. Prev.* **43**, 533–548 (2011).
3. Iranzo, A. & Santamaria, J. Sleep in neurodegenerative diseases. *Sleep Medicine: A Comprehensive Guide to Its Development, Clinical Milestones, and Advances in Treatment* 271–283 (Springer, 2015).
4. Tsuno, N., Besset, A. & Ritchie, K. et al. Sleep and depression. *J. Clin. psychiatry* **66**, 1254–1269 (2005).
5. Martin, S. E., Engleman, H. M., Deary, I. J. & Douglas, N. J. The effect of sleep fragmentation on daytime function. *Am. J. Respir. Crit. Care Med.* **153**, 1328–1332 (1996).
6. Strollo Jr, P. J. & Rogers, R. M. Obstructive sleep apnea. *N. Engl. J. Med.* **334**, 99–104 (1996).
7. Boeve, B. F. Idiopathic rem sleep behaviour disorder in the development of parkinson’s disease. *Lancet Neurol.* **12**, 469–482 (2013).
8. Stephansen, J. B. et al. Neural network analysis of sleep stages enables efficient diagnosis of narcolepsy. *Nat. Commun.* **9**, 5229 (2018).
9. Bohnen, N. I. & Hu, M. Sleep disturbance as potential risk and progression factor for parkinson’s disease. *J. Parkinson’s. Dis.* **9**, 603–614 (2019).
10. Berry, R. B. et al. The aasm manual for the scoring of sleep and associated events. *Rules, Terminol. Tech. Specif., Darien, Ill., Am. Acad. Sleep. Med.* **176**, 7 (2012).
11. Magalang, U. J. et al. Agreement in the scoring of respiratory events and sleep among international sleep centers. *Sleep* **36**, 591–596 (2013).
12. Younes, M., Raneri, J. & Hanly, P. Staging sleep in polysomnograms: analysis of inter-scoring variability. *J. Clin. Sleep. Med.* **12**, 885–894 (2016).
13. Biswal, S. et al. Sleepnet: automated sleep staging system via deep learning. arXiv preprint arXiv:1707.08262 (2017).

14. Biswal, S. et al. Expert-level sleep scoring with deep neural networks. *J. Am. Med. Inform. Assoc.* **25**, 1643–1650 (2018).
15. Perslev, M., Jensen, M., Darkner, S., Jennum, P. J. & Igel, C. U-time: a fully convolutional network for time series segmentation applied to sleep staging. *Adv. Neural Inform. Process. Syst.* **32** (2019).
16. Sridhar, N. et al. Deep learning for automated sleep staging using instantaneous heart rate. *NPJ Digital Med.* **3**, 106 (2020).
17. Tveit, J. et al. Automated interpretation of clinical electroencephalograms using artificial intelligence. *JAMA Neurol.* **80**, 805–812 (2023).
18. Uchida, S., Maloney, T., March, J., Azari, R. & Feinberg, I. Sigma (12–15 Hz) and delta (0.3–3 Hz) EEG oscillate reciprocally within NREM sleep. *Brain Res. Bull.* **27**, 93–96 (1991).
19. Uchida, S., Maloney, T. & Feinberg, I. Beta (20–28 Hz) and delta (0.3–3 Hz) EEGs oscillate reciprocally across NREM and REM sleep. *Sleep* **15**, 352–358 (1992).
20. Bonnet, M. & Arand, D. Heart rate variability: sleep stage, time of night, and arousal influences. *Electroencephalogr. Clin. Neurophysiol.* **102**, 390–396 (1997).
21. Johnson, L., Church, M., Seales, D. & Rossiter, V. Auditory arousal thresholds of good sleepers and poor sleepers with and without flurazepam. *Sleep* **1**, 259–270 (1978).
22. Gleeson, K., ZWILLICH, C. W. & White, D. P. The influence of increasing ventilatory effort on arousal from sleep 1–3. *Am. Rev. Respir. Dis.* **142**, 295–300 (1990).
23. Younes, M. et al. Odds ratio product of sleep EEG as a continuous measure of sleep state. *Sleep* **38**, 641–654 (2015).
24. Meza-Vargas, S., Giannouli, E. & Younes, M. Enhancements to the multiple sleep latency test. *Nat. Sci. Sleep* **8**, 145–158 (2016).
25. Qanash, S., Giannouli, E. & Younes, M. Assessment of intervention-related changes in non-rapid-eye-movement sleep depth: importance of sleep depth changes within stage 2. *Sleep. Med.* **40**, 84–93 (2017).
26. Younes, M., Soiferman, M., Thompson, W. & Giannouli, E. Performance of a new portable wireless sleep monitor. *J. Clin. Sleep. Med.* **13**, 245–258 (2017).
27. Younes, M., Azarbarzin, A., Reid, M., Mazzotti, D. R. & Redline, S. Characteristics and reproducibility of novel sleep EEG biomarkers and their variation with sleep apnea and insomnia in a large community-based cohort. *Sleep* **44**, zsab145 (2021).
28. Younes, M. et al. Sleep architecture based on sleep depth and propensity: patterns in different demographics and sleep disorders and association with health outcomes. *Sleep* **45**, zsa059 (2022).
29. Metzner, C. et al. Extracting continuous sleep depth from EEG data without machine learning. *Neurobiol. Sleep. Circadian Rhythms* **14**, 100097 (2023).
30. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
31. Goldstein, C. A. et al. Artificial intelligence in sleep medicine: an American Academy of Sleep Medicine position statement. *J. Clin. Sleep. Med.* **16**, 605–607 (2020).
32. Bandyopadhyay, A. & Goldstein, C. Clinical applications of artificial intelligence in sleep medicine: a sleep clinician's perspective. *Sleep. Breath.* **27**, 39–55 (2023).
33. Liu, T.-Y. et al. Learning to rank for information retrieval. *Found. Trends® Inf. Retr.* **3**, 225–331 (2009).
34. Karatzoglou, A., Baltrunas, L. & Shi, Y. Learning to rank for recommender systems. *Proc. 7th ACM Conference on Recommender Systems* 493–494 (Association for Computing Machinery, 2013).
35. Agarwal, A. et al. Learning to rank for robust question answering. *Proc. 21st ACM international conference on Information and knowledge management* 833–842 (Association for Computing Machinery, 2012).
36. Patel, A. K., Reddy, V., Shumway, K. R. & Araujo, J. F. Physiology, sleep stages. In StatPearls [Internet] (StatPearls Publishing, 2022).
37. Ferini-Strambi, L. & Zucconi, M. REM sleep behavior disorder. *Clin. Neurophysiol.* **111**, S136–S140 (2000).
38. Postuma, R., Gagnon, J.-F., Rompre, S. & Montplaisir, J. Severity of REM atonia loss in idiopathic REM sleep behavior disorder predicts Parkinson disease. *Neurology* **74**, 239–244 (2010).
39. McCarter, S. J., St. Louis, E. K. & Boeve, B. F. REM sleep behavior disorder and REM sleep without atonia as an early manifestation of degenerative neurological disease. *Curr. Neurol. Neurosci. Rep.* **12**, 182–192 (2012).
40. Scammell, T. E., Arrigoni, E. & Lipton, J. O. Neural circuitry of wakefulness and sleep. *Neuron* **93**, 747–765 (2017).
41. Kleiger, R. E., Miller, J. P., Bigger Jr, J. T. & Moss, A. J. Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. *Am. J. Cardiol.* **59**, 256–262 (1987).
42. Bigger Jr, J. T. et al. Frequency domain measures of heart period variability and mortality after myocardial infarction. *Circulation* **85**, 164–171 (1992).
43. Wolpert, E. A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. *Arch. Gen. Psychiatry* **20**, 246–247 (1969).
44. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
45. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
46. Van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat. Med.* **27**, 775–784 (2021).
47. Courtiol, P. et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nat. Med.* **25**, 1519–1525 (2019).
48. Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
49. Liu, Y. et al. A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* **26**, 900–908 (2020).
50. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inform. Process. Syst.* **30**, 6000–6010 (2017).
51. Kaplan, J. et al. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020).
52. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. (ICLR, 2021).
53. Yang, C., Westover, M. & Sun, J. Biot: Biosignal transformer for cross-data learning in the wild. *Adv. Neural Inform. Process. Syst.* **36**, 78240–78260 (2024).
54. Boe, A. J. et al. Automating sleep stage classification using wireless, wearable sensors. *NPJ Digital Med.* **2**, 131 (2019).
55. Radha, M. et al. A deep transfer learning approach for wearable sleep stage classification with photoplethysmography. *NPJ Digital Med.* **4**, 135 (2021).
56. Perslev, M. et al. U-sleep: resilient high-frequency sleep staging. *NPJ Digital Med.* **4**, 72 (2021).
57. Perslev, M. et al. Automatic detection of abnormal sleeping patterns in stroke patients using high-frequency sleep staging. *J. Sleep Res.* **31** (2022).
58. Zhang, G.-Q. et al. The national sleep research resource: towards a sleep data commons. *J. Am. Med. Inform. Assoc.* **25**, 1351–1358 (2018).
59. Quan, S. F. et al. The sleep heart health study: design, rationale, and methods. *Sleep* **20**, 1077–1085 (1997).
60. Redline, S. et al. The familial aggregation of obstructive sleep apnea. *Am. J. Respir. Crit. Care Med.* **151**, 682–687 (1995).
61. Chen, X. et al. Racial/ethnic differences in sleep disturbances: the multi-ethnic study of atherosclerosis (MESA). *Sleep* **38**, 877–888 (2015).
62. 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).
63. Gramfort, A. et al. Meg and eeg data analysis with mne-python. *Front. Neurosci.* **7**, 70133 (2013).
64. Punjabi, N. M. et al. Sleep-disordered breathing and mortality: a prospective cohort study. *PLoS Med.* **6**, e1000132 (2009).
65. Richman, J. S. & Moorman, J. R. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* **278**, H2039–H2049 (2000).
66. Hardstone, R. et al. Detrended fluctuation analysis: a scale-free view on neuronal oscillations. *Front. Physiol.* **3**, 450 (2012).
67. Bao, Y., Sivanandan, S. & Karaletsos, T. Channel vision transformers: an image is worth 1x16x16 words. (ICLR, 2024).
68. Ba, J. L. Layer normalization. arXiv preprint arXiv:1607.06450 (2016).
69. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proc. IEEE conference on computer vision and pattern recognition* 770–778 (Institute of Electrical and Electronics Engineers, 2016).

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (62102008, 62172018); Clinical Medicine Plus X - Young Scholars Project of Peking University, the Fundamental Research Funds for the Central Universities (PKU2024LCXQ030); PKU-OPPO Fund (B0202301); CCF-Zhipu Large Model Innovation Fund (CCF-Zhipu202414), the Ministry of Science and Technology of the People's Republic of China (STI2030-Major Projects2021ZD0201900). Dr. Westover's laboratory received support from grants from the NIH (R01NS102190, R01NS102574, R01NS107291, RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598) and NSF (2014431).

## Author contributions

S.Z.: conceptualization, data curation, methodology, formal analysis, validation, and manuscript writing. G.S.: data curation, methodology, formal analysis, and manuscript writing. H.S.: conceptualization, formal analysis, validation, and manuscript writing. D.Z.: data curation, software, and validation. Y.L.: conceptualization, validation, manuscript writing, and supervision. M.B.W.: conceptualization, validation, funding acquisition, and supervision. S.H.: conceptualization, methodology, formal analysis, funding

acquisition, manuscript writing, and supervision. All authors have read and approved the manuscript.

## Competing interests

M.B.W. is a co-founder, scientific advisor, consultant to, and has personal equity interest in Beacon Biosignals. The other authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41746-025-01607-0>.

**Correspondence** and requests for materials should be addressed to Shenda Hong.

**Reprints and permissions 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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence 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 licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025