A Zero-Burden Sleep Foundation Model Built on Cardiorespiratory Signals from 800,000+ Hours of Multi-Ethnic Sleep Recordings

Guangkun Nie, *B.S.*,[1,2,#] Xuesong Chen, *M.S.*,[3,#] Yichen Wang, *M.S.*,[3,#] Jingxu Chen, *M.D.*,[4] Yunhan Shi, *M.D.*,[5] Jianwen Zhong, *M.D.*,[6] Jie Shi, *Ph.D.*,[7] Chun-feng Liu, *M.D.*,[8] Bei Huang, *M.D.*,[9,10] Yaping Liu *M.D.*,[9,11] Jihui Zhang, *M.D.*,[9,11] Yi Fang, *M.D.*,[12] Haoqi Sun, *Ph.D.*,[13] Robert J. Thomas, *M.D.*,[14] Weijun Huang, *M.D.*,[15] Zengrui Jin, *Ph.D.*,[16] Fei Lei, *M.S.*,[17] Leilei Wang, *M.D.*,[4] Rui Zhao, *M.D.*,[18] Chao Zhang, *Ph.D.*,[16] Kaibing Chen, *M.D.*,[19] Dongsheng Lv, *M.D.*,[18] Wei Chen, *M.D.*,[20] Hongliang Yi, *M.D.*,[15] Jun Liu, *M.D.*,[21] Yun-Kwok Wing, *M.D.*,[9,10] Hongyan Li, *Ph.D.*,[2,22] M. Brandon Westover, *M.D.*, *Ph.D.*,[13] Lin Lu, *M.D.*,[7] Xiangdong Tang, *M.D.*,[17] Shankai Yin, *M.D.*, *Ph.D.*,[15] Yanru Li, *M.D.*,[5,*] Shenda Hong, *Ph.D.*,[1,*] Yue Leng, *Ph.D.*[12,*] on behalf of the AISleep 365 Consortium

[1] National Institute of Health Data Science, Peking University, Beijing, China

[2] School of Intelligence Science and Technology, Peking University, Beijing, China

[3] Five Seasons Medical, Beijing, China

[4] Sleep Medicine Center, Beijing HuiLongGuan Hospital, Peking University, Beijing, China

[5] Department of Otorhinolaryngology Head and Neck Surgery, Beijing Tongren Hospital, Capital Medical University, Beijing, China

[6] Department of Pediatric Otorhinolaryngology, Shenzhen Hospital of Southern Medical University, Shenzhen, China

[7] National Institute on Drug Dependence and Beijing Key Laboratory of Drug Dependence Research, Peking University, Beijing, China

[8] Department of Neurology and Clinical Research Center of Neurological Disease, Second Affiliated Hospital of Soochow University, Suzhou, Jiangsu, China

[9] Li Chiu Kong Family Sleep Assessment Unit, Department of Psychiatry, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China

[10] Li Ka Shing Institute of Health Sciences, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China

[11] Center for Sleep and Circadian Medicine, Affiliated Brain Hospital of Guangzhou Medical University, Guangzhou, Guangdong, China

[12] Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, CA, USA

[13] Department of Neurology, Beth Israel Deaconess Medical Center, Boston, MA, USA

[14] Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, United States

[15] Department of Otorhinolaryngology Head and Neck Surgery, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China

[16] Department of Electronic Engineering, Tsinghua University, Beijing, China

[17] Sleep Medicine Center, West China Hospital, Sichuan University, Chengdu, Sichuan, China

[18] Sleep Medicine Center, Inner Mongolia Mental Health Center (The Third Hospital of Inner Mongolia

Autonomous Region, Brain Hospital of Inner Mongolia Autonomous Region), Hohhot, Inner Mongolia, China

[19] Sleep Medicine Center, Affiliated Hospital of Gansu University of Chinese Medicine, Lanzhou, Gansu, China

[20] Department of Psychiatry, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China

[21] Department of Neurology and Institute of Neurology, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[22] State Key Laboratory of General Artificial Intelligence, Peking University, Beijing, China

# These authors contributed equally to this work.

* Co-Corresponding authors

Yue Leng: yue.leng@ucsf.edu

Shenda Hong: hongshenda@pku.edu.cn

Yanru Li: liyanruru@aliyun.com

Abstract

BACKGROUND Sleep disorders pose a major global health burden and are associated with a wide range of adverse health outcomes. Polysomnography (PSG) is the gold standard for sleep assessment, but it is impractical for home-based or long-term monitoring. We investigated whether zero-burden cardiorespiratory signals, when harnessed through foundation model approaches, can enable accurate sleep assessment at scale and capture broader dimensions of multi-organ health.

METHODS We present SleepFounder, a foundation model for zero-burden sleep monitoring built upon cardiorespiratory signals. SleepFounder was developed on the largest curated multi-ethnic sleep dataset to date, comprising over 800,000 hours of recordings from 35 cohorts across the United States and China. We evaluated SleepFounder across downstream tasks ranging from conventional sleep analysis to emerging applications, including demographic profiling and multi-organ disease detection and prediction. We further conducted a real-world study using multi-center ballistocardiography (BCG) data collected with a custom-developed sleep mat system for external validation.

RESULTS SleepFounder achieved strong performance across diverse downstream tasks and consistently outperformed baseline models, obtaining the best results in 14 out of 17 dataset-task pairs. For conventional sleep analysis and demographic profiling, averaged across external datasets, it achieved a Cohen's Kappa of 0.671 (0.668-0.673) for five-class sleep staging, an area under the receiver operating characteristic curve (AUROC) of 0.917 (0.912-0.922) for moderate-to-severe obstructive sleep apnea detection, a mean absolute error of 6.727 (6.684-6.771) years for age prediction, and an AUROC of 0.865 (0.860-0.870) for sex classification. In multi-organ disease detection, representative AUROCs reached 0.943 (0.917-0.966) for Parkinson's disease, 0.886 (0.841-0.928) for gastroesophageal reflux disease, and 0.881 (0.831-0.922) for heart failure. Additional conditions, including high cholesterol, coronary heart disease (CHD), bipolar disorder, and chronic pain, achieved AUROCs ranging from 0.811 to 0.830 in the held-out test set, with results further validated across five external cohorts. For future disease prediction, concordance indices reached 0.838 (0.797-0.873) for

CHD death and 0.837 (0.806-0.865) for cardiovascular disease death, with corresponding metrics of 0.734-0.781 for congestive heart failure, stroke, and angina. In the real-world BCG study, SleepFounder maintained 94% of its performance on average relative to prior external validations conducted on PSG-based datasets.

CONCLUSIONS SleepFounder establishes a foundation model that learns from cardiorespiratory signals to enable accurate, scalable, and zero-burden sleep assessment. By linking sleep physiology with multi-organ health, it bridges clinical and home settings and demonstrates that signals traditionally used for sleep monitoring can serve as powerful biomarkers of systemic function and disease risk. These findings highlight a new paradigm for zero-burden sleep and health monitoring in real-world settings.

Introduction

Sleep is essential for overall health, with poor sleep quality linked to increased risks of cardiovascular diseases (CVDs)[1,2,3,4], metabolic dysfunction[5,6], neurological disorders[7,8], and psychiatric conditions[9,10]. Sleep disorders, such as obstructive sleep apnea (OSA), are a growing global health concern. For example, an estimated 936 million adults aged 30-69 worldwide suffer from mild to severe sleep apnea[11]. In the United States, 80% of OSA cases among nearly 85 million affected individuals remain undiagnosed[12,13]. Polysomnography (PSG), the gold standard for sleep analysis, captures high-fidelity physiological signals such as electroencephalography (EEG), respiration, electrocardiography (ECG), and blood oxygen saturation (SpO_2). Despite its accuracy, PSG is resource-intensive and disrupts natural sleep, making it unsuitable for home-based or long-term monitoring[14,15]. There is an urgent need for cost-effective, zero-burden solutions to support reliable sleep monitoring.

Recent studies[16,17,18] have explored the use of reduced subsets of PSG channels to enable zero-burden sleep monitoring. In particular, deep learning methods applied to cardiorespiratory signals have shown promising performance in key sleep analysis tasks, including sleep staging[19,20,21,22] and OSA severity/apnea-hypopnea index (AHI) estimation[23,24,25,26]. These signals, which capture the dynamics of the cardiovascular and respiratory systems, are strongly linked to sleep patterns and overall physiological states. Importantly, they represent a shared physiological modality captured in both clinical PSG and home-based systems such as ballistocardiography (BCG)[22,26], creating the opportunity to assemble large, heterogeneous datasets well suited for advanced deep learning techniques. However, most prior studies have overlooked this shared characteristic, relying on limited dataset size and demographic diversity, thereby constraining the accuracy, generalizability, and robustness of existing approaches. Foundation models, which leverage large-scale pretraining and generalize across diverse downstream tasks, have shown considerable promise in the medical domain, including computational pathology[27,28], echocardiography[29], and physiological signal analysis[30,31,32,33,34]. Building on these advances, we propose that large-scale, physiologically salient cardiorespiratory data—derived from both PSG and home-based devices—combined with foundation model techniques, can enable accurate, zero-burden sleep assessment and extend toward broader multi-organ health evaluation.

In this study, we introduce SleepFounder, a foundation model for accurate, zero-burden sleep monitoring based solely on cardiorespiratory signals (**Figure 1**). The model was developed on the largest curated multi-ethnic sleep dataset to date, totaling over 800,000 hours of sleep recordings from

35 cohorts across the United States and China. During pretraining, SleepFounder was guided by two physiology-inspired pretext tasks, EEG spectrogram reconstruction and oxygen desaturation prediction, which enabled the model to extract physiologically meaningful information embedded in cardiorespiratory signals and to internalize latent mechanisms linking these dynamics to sleep. Unlike prior foundation models such as SleepFM[34], which rely on resource-intensive PSG data and cannot be readily adapted to home-based sleep monitoring, SleepFounder builds on widely accessible cardiorespiratory signals, enabling flexible, zero-burden sleep analysis with potential for population-scale deployment.

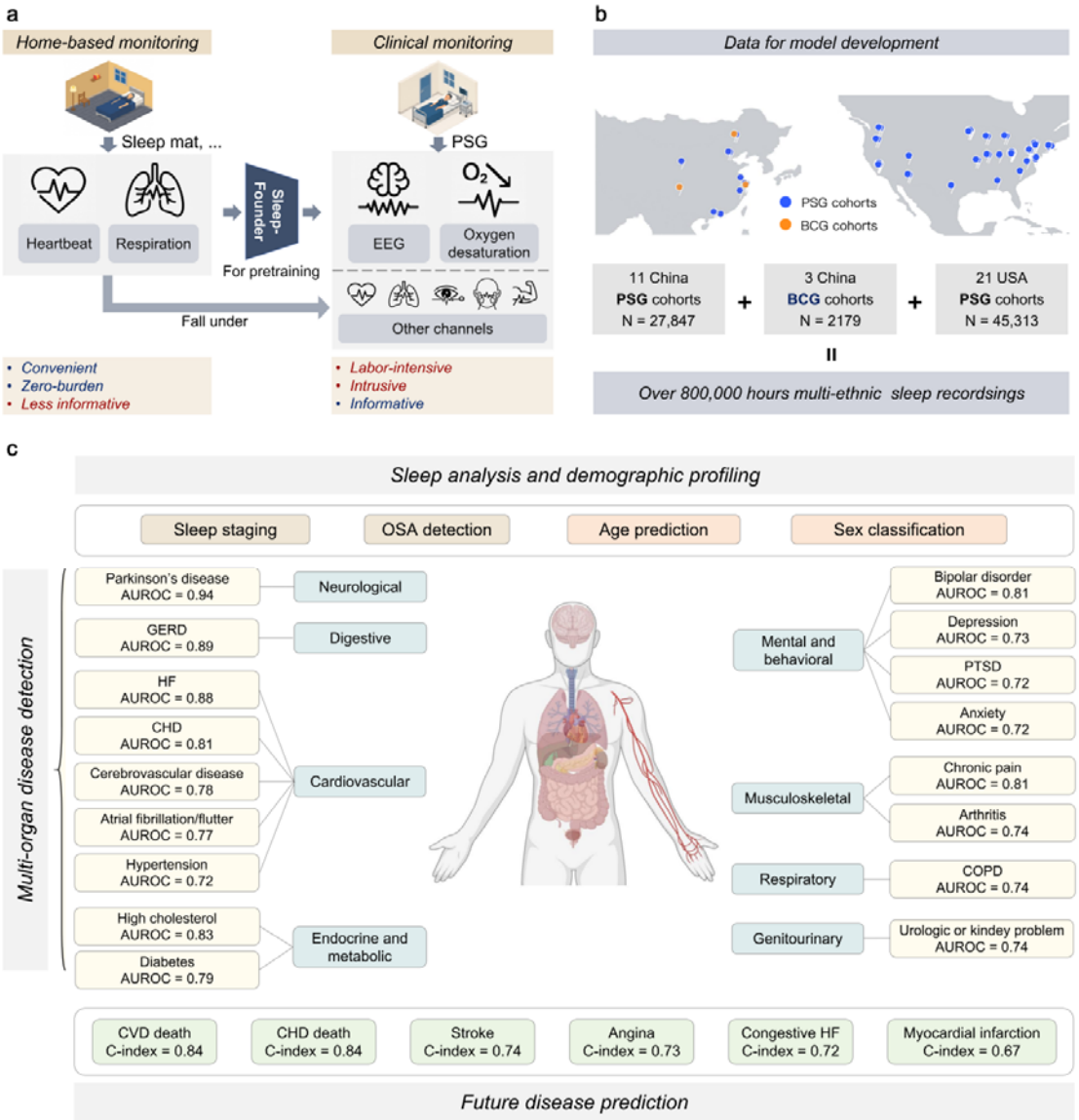


Figure 1. Overview of the proposed SleepFounder model. (a) SleepFounder leverages respiratory and heartbeat signals shared between clinical PSG systems and contactless home-based sleep monitoring devices. During pretraining, the model learns to reconstruct EEG and oxygen desaturation signals, using informative modalities to guide the representation learning of less informative ones. (b) SleepFounder was developed using large-scale, multi-ethnic sleep datasets from China and the United States, comprising eleven Chinese PSG

cohorts, three real-world BCG cohorts, and twenty-one USA PSG cohorts, totaling over 800,000 hours of recorded sleep data. (c) SleepFounder was evaluated across multiple downstream tasks, including sleep analysis, demographic profiling, multi-organ disease detection, and future disease prediction. The results shown correspond to internal test datasets. PSG, polysomnography; EEG, electroencephalography; BCG, ballistocardiography. OSA, obstructive sleep apnea; GERD, gastroesophageal reflux disease; HF, heart failure; CHD, coronary heart disease; PTSD, post-traumatic stress disorder; COPD, chronic obstructive pulmonary disease; CVD, cardiovascular disease.

Methods

Study design and oversight

This study used both public and private datasets and was conducted in accordance with the principles of the Declaration of Helsinki. Ethical approval for the overall project was obtained from the Biomedical Ethics Review Committee of Peking University (IRB00001052-23207). In addition, local institutional review board approvals were obtained from all participating clinical centers, including the Affiliated Hospital of Gansu University of Chinese Medicine (AHGUCM), Beijing Huilongguan Hospital (BHH), Inner Mongolia Mental Health Center (IMMHC), Shenzhen Hospital of Southern Medical University (SHSMU), Beijing Tongren Hospital (BTH affiliated to Capital Medical University), Sir Run Run Shaw Hospital, (SRRSH affiliated to Zhejiang University School of Medicine), Shanghai Sixth People's Hospital (SSPH affiliated to Shanghai Jiao Tong University School of Medicine), the Sleep Medicine Center of West China Hospital (WCH affiliated to Sichuan University), the Second Affiliated Hospital of Soochow University (SAHSU), Ruijin Hospital (RH affiliated to Shanghai Jiao Tong University School of Medicine), the Affiliated Brain Hospital of Guangzhou Medical University (ABHGMU), and the Chinese University of Hong Kong Medical Center (CUHKMC). As only de-identified retrospective data were used, the requirement for informed consent was waived by each institutional review board.

Pretraining datasets for SleepFounder

We curated a large-scale, multi-ethnic dataset for pretraining SleepFounder, comprising PSG data from 45,565 individuals across 24 cohorts in the United States and China. The U.S. component included 19 cohorts with 30,913 participants, encompassing the Human Sleep Project (HSP) and 18 datasets from the National Sleep Research Resource (NSRR). The Chinese component consisted of five private clinical cohorts with 14,652 participants, namely AHGUCM, BHH, IMMHC, SHSMU, and BTH. Detailed descriptions of these cohorts are provided in Supplementary Methods, Section 1.1.

Data preprocessing for pretraining

Modality-specific procedures were applied to extract shared cardiorespiratory features between PSG and BCG recordings, specifically heartbeat and respiratory signals. Overnight EEG and SpO_2 data were further converted into spectrograms and desaturation curves, respectively, which served as reconstruction targets during pretraining. Detailed preprocessing steps are provided in Supplementary Methods, Section 2.

Model architecture and pretraining details

SleepFounder consists of a ResNet-based[35] feature extractor followed by an RoFormer encoder[36]. The model takes overnight heartbeat and respiratory signals as input and adopts an early fusion strategy to facilitate cross-modal feature interaction. During feature extraction, two structurally identical but independently parameterized ResNet branches extract temporal features from each physiological modality. Each branch begins with a preprocessing module comprising a convolutional layer, batch normalization, ReLU activation, and max-pooling, followed by eight residual blocks that progressively transform raw signals into 512-channel local representations. Feature fusion is achieved through element-wise summation, enabling the learning of shared representations. To capture global temporal dependencies, SleepFounder integrates an RoFormer encoder. This encoder comprises six stacked layers, each with eight attention heads, a hidden dimension of 512, and a feedforward dimension of 2048. This hierarchical architecture allows SleepFounder to jointly learn fine-grained physiological patterns through the ResNet encoders while leveraging the RoFormer to capture long-range dependencies across sleep stages.

The pretraining stage is designed to enable the model to learn semantically meaningful and physiologically grounded representations by jointly optimizing two pretext tasks: EEG spectrogram reconstruction and oxygen desaturation prediction. These tasks capture complementary aspects of sleep physiology, with the former focusing on neural dynamics and the latter on cardiorespiratory patterns. Features extracted by the RoFormer encoder are passed through two task-specific linear heads. The EEG reconstruction head is trained using mean squared error (MSE) between predicted and ground-truth spectrograms. The oxygen desaturation head predicts continuous SpO_2 trajectories over 30-second intervals, with a loss combining Wasserstein distance, which captures cumulative deviations over time, and Pearson correlation loss, which promotes linear alignment with ground-truth trends. The combined SpO_2 loss is defined as $\mathcal{L}_{\mathrm{SpO2}} = \lambda_{\mathrm{Pearson}}\mathcal{L}_{\mathrm{Pearson}} + \lambda_{\mathrm{Wasserstein}}\mathcal{L}_{\mathrm{Wasserstein}}$, where $\lambda_{\mathrm{Pearson}}$ and $\lambda_{\mathrm{Wasserstein}}$ were set to 0.1 and 0.9, respectively. The overall pretraining objective integrates both tasks as a weighted sum, $\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{EEG}} + \lambda_{\mathrm{SpO2}}\mathcal{L}_{\mathrm{SpO2}}$, with the weight λ_{SpO2} set to 0.1.

During pretraining, data were randomly partitioned at the participant level into training, validation, and test sets with an 8:1:1 ratio, maintaining consistent proportions within each individual dataset. An exception was made for the HSP dataset, which was split temporally: recordings collected before July 1, 2016 were used for training and validation in an 8:1 ratio, while those collected afterward were reserved exclusively for temporally external validation. The model was trained on four NVIDIA H20 GPUs with a batch size of 65 and an initial learning rate of 1e-4, using the Adam optimizer. Training proceeded for up to 20 epochs with early stopping: if the validation loss did not improve for 10 consecutive epochs, training was terminated to prevent overfitting.

Fine-tuning strategy

To adapt SleepFounder for downstream tasks, the EEG spectrogram reconstruction and oxygen desaturation prediction heads were removed and replaced with a single randomly initialized linear layer appended to the RoFormer encoder. Rather than relying solely on the final layer's hidden states, representations from all encoder layers are aggregated using learnable weights. Specifically, the aggregated hidden state is computed as $\mathbf{h}^{\mathrm{agg}} = \sum_{l=1}^{L} \alpha_l \mathbf{h}^{(l)}$, where $\mathbf{h}^{(l)}$ denotes the hidden state from

layer l, L is the total number of layers, and α_l are learnable aggregation weights initialized such that $\alpha_L = 1$ and $\alpha_{1:L-1} = 0$. These weights are optimized during fine-tuning to prioritize task-relevant layers.

Fine-tuning was conducted for 10 epochs using the same optimizer and hyperparameters as in pretraining. Participant-level data were generally partitioned into training, validation, and testing sets at an 8:1:1 ratio. External test datasets were evaluated in their entirety without further partitioning. All experiments were performed on eight NVIDIA H20 GPUs.

Downstream Evaluation Details

Datasets for downstream evaluation

To comprehensively evaluate SleepFounder, we utilized both well-annotated PSG datasets and real-world, multi-center BCG datasets collected using our previously developed Five Seasons (5S) sleep tracking mat[37]. The PSG evaluation datasets included the temporally split HSP dataset and the two publicly available NSRR cohorts excluded from pretraining, namely the Sleep Heart Health Study (SHHS) and the Wisconsin Sleep Cohort (WSC), as well as two clinical datasets, SRRSH and SSPH, from China. For PD detection, we specifically curated a multi-source dataset, in which cases were drawn from multiple clinics in China, including SAHSU, RH, and ABHGMU, and controls were collected from CUHKMC, SHHS, the Heart Biomarker Evaluation in Apnea Treatment (HeartBEAT), the Best Apnea Interventions in Research (BestAIR), and the Home Positive Airway Pressure (HomePAP) datasets. The multi-center BCG datasets were collected from SSPH, WCH, and IMMHC, in parallel with clinical PSG recordings. Further details of these datasets are provided in the Supplementary Methods, Section 1.2.

Baseline Methods for comparison

To evaluate the performance of SleepFounder on downstream tasks, we established a set of baseline models, including the self-supervised pretrained model SleepFM[34] and conventional task-specific models for sleep staging and moderate-to-severe OSA detection. For fair comparison, all baselines were adapted to the unified 4 Hz respiratory and heartbeat signals described above. Specifically, because SleepFM was not originally designed for home-based sleep monitoring but instead relied on a contrastive learning approach to align representations across PSG channels during pretraining, including EEG, ECG, EMG, and respiratory signals, we adapted SleepFM (referred to as SleepFM-CR) using the same pretraining framework together with our larger and more diverse pretraining dataset. This adaptation was performed to demonstrate the improvement achieved by our physiology-inspired pretraining method compared with the contrastive learning approach.

Task design and evaluation metrics

We evaluated SleepFounder across six categories of downstream tasks, encompassing conventional sleep analysis (five-class sleep staging and moderate-to-severe OSA detection, defined as AHI≥15 and hereafter referred to as OSA detection for simplicity), demographic profiling (age prediction and sex classification), multi-organ disease detection, and future disease prediction. For each task, the model

produced outputs that directly corresponded to the respective prediction target. The only exception was OSA detection, where the model first estimated respiratory event probabilities at a 1-second resolution. A probability threshold optimized on the validation set was then applied to derive binary event labels, and consecutive abnormal segments were merged into apnea or hypopnea events according to their duration. From these detected events, the AHI was computed as the number of apnea or hypopnea events per hour of sleep. In this task, the model's continuous AHI predictions were further converted into binary labels to determine the presence of OSA.

Task-specific evaluation metrics were selected according to the prediction objective. For five-class sleep staging, performance was assessed using Cohen's kappa coefficient, which quantifies the agreement between predicted and reference labels while accounting for chance. For OSA detection, evaluation was performed using the AUROC (calculated based on the ranking of the predicted AHI values), where a value of 1 indicates perfect discrimination and 0.5 represents random chance. For demographic profiling, age prediction was assessed using the mean absolute error (MAE), representing the average absolute deviation from the true age. Sex classification and multi-organ disease detection were evaluated using the AUROC. For future disease prediction, model performance was quantified using the 6-year AUROC and concordance index (C-index), which jointly capture discriminative accuracy and time-to-event concordance.

Results

Study Population and Method Overview

SleepFounder was developed using large-scale, multi-center, and demographically diverse datasets from the United States and China (35 in total). Pretraining involved 24 datasets comprising 45,565 individuals and totaling 489,481 hours of sleep recordings. Model evaluation was conducted on four independent PSG cohorts (SHHS, WSC, SSPH, and SRRSH), and real-world validation was performed using multi-center BCG cohorts (n = 2,179) collected with the zero-burden 5S sleep tracking mat. Baseline demographic characteristics of all cohorts are summarized in **Table 1**. Because race classification schemes differed across datasets, we adopted a universal superset of race categories to present each dataset's original taxonomy.

Table 1. Summary statistics of datasets used for pretraining and evaluation.

Dataset	No. subjects (Recordings)	Age	Male	Race/ Ethnicity				
				White	Asian	Black or African American	Hispanic	Others
ABC	48 (125)	50.1 ± 18.8	78 (59.1%)	95 (77.2%)	4 (3.3%)	24 (19.5%)	0 (0%)	0 (0%)
ApoE	712 (712)	45.7 ± 27.2	421 (59.1%)	–	–	–	–	—
APPLES	1095 (1095)	51.6 ± 24.5	719 (65.7%)	835 (76.3%)	61 (5.6%)	99 (9.0%)	77 (7.03%)	23 (2.1%)
BestAIR	428 (514)	64.8 ± 14.2	170 (67.7%)	229 (92.0%)	6 (2.4%)	14 (5.6%)	–	0 (0%)
CCSHS	501 (501)	17.7 ± 0.9	256 (51.1%)	299 (59.7%)	0 (0%)	180 (35.9%)	–	22 (4.4%)
CFS	730 (730)	41.4 ± 38.8	329 (45.1%)	302 (41.4%)	–	406 (55.6%)	–	22 (3.0%)
CHAT	1006 (1295)	7.0 ± 2.9	327 (48.9%)	253 (37.8%)	–	331 (49.5%)	–	85 (12.7%)
HeartBEAT	316 (583)	63.2 ± 14.5	428 (73.4%)	472 (81.2%)	12 (2.1%)	75 (12.9%)	–	22 (3.8%)
HomePAP	343 (414)	46.4 ± 23.6	239 (57.7%)	298 (73.6%)	5 (1.2%)	90 (22.2%)	–	12 (3.0%)
HSP	18702 (23290)	53.0 ± 33.6	13323 (57.2%)	–	–	–	–	—
MESA	2056 (2056)	69.3 ± 18.2	954 (46.4%)	743 (36.1%)	250 (12.2%)	572 (27.8%)	491 (23.9%)	—
MNC	982 (982)	–	–	–	–	–	–	—
MrOS	2905 (3930)	77.6 ± 11.2	3930 (100%)	3540 (92.6%)	134 (3.5%)	140 (3.7%)	6 (0.2%)	2 (0.1%)
MSP	106 (106)	26.8 ± 11.9	0 (0%)	8 (7.6%)	2 (1.9%)	95 (90.5%)	–	—
NCHSDB	3644 (3952)	8.8 ± 11.9	2247 (56.9%)	–	–	–	–	—
nuMoM2b	3141 (5240)	27.0 ± 11.0	0 (0%)	3269 (62.4%)	192 (3.7%)	592 (11.3%)	945 (18.0%)	242 (4.6%)
PATS	383 (547)	7.1 ± 4.7	290 (53.0%)	374 (68.9%)	11 (2.0%)	126 (23.2%)	–	32 (5.9%)
SOF	453 (453)	82.9 ± 6.3	0 (0%)	415 (91.6%)	–	38 (8.4%)	–	—

(Table continued on next page)

Table 1 (continued). Summary statistics of datasets used for pretraining and evaluation.

Dataset	**No. subjects (Recordings)**	**Age**	**Male**	**Race/ Ethnicity**				
				White	Asian	Black or African American	Hispanic	Others
STAGES	844 (848)	44.7 ± 29.3	414 (52.3%)	555 (70.5%)	113 (14.4%)	41 (5.2%)	–	78 (10.0%)
AHGUCM	1914 (1914)	46.7 ± 32.1	1250 (72.6%)	–	1914 (100%)	–	–	—
BHH	908 (950)	39.7 ± 39.2	125 (43.1%)	–	908 (100%)	–	–	—
IMMHC	2046 (2046)	52.0 ± 30.6	704 (35.4%)	–	2046 (100%)	–	–	—
SHSMU	2854 (2854)	8.0 ± 6.8	1667 (66.0%)	–	2854 (100%)	–	–	—
BTH	6930 (6930)	34.5 ± 17.8	5357 (77.3%)	–	6930 (100%)	–	–	—
SHHS	5795 (8440)	64.5 ± 22.3	3983 (47.2%)	7206 (85.4%)	–	695 (8.2%)	–	
WSC	1123 (2570)	59.8 ± 17.0	1385 (53.9%)	2463 (95.8%)	23 (0.9%)	36 (1.4%)	26 (1.0%)	539 (6.39%)
SSPH	7429 (7429)	40.0 ± 25.5	5814 (78.3%)	–	7429 (100%)	–	–	22 (0.9%)
SRRSH	4384 (4611)	50.1 ± 31.7	1970 (42.8%)	–	4611 (100%)	–	–	—
RH	161 (161)	65.2 ± 19.0	56 (51.4%)	–	161 (100%)	–	–	—
SAHSU	130 (130)	61.4 ± 37.2	83 (66.9%)	–	130 (100%)	–	–	—
ABHGMU	147 (147)	57.8 ± 20.3	77 (52.4%)	–	147 (100%)	–	–	—
CUHKMC	265 (265)	68.4 ± 25.2	131 (49.4%)	–	265 (100%)	–	–	—
BCG_{WCH}	945 (945)	42.1 ± 31.1	693 (73.3%)	–	945 (100%)	–	–	—
BCG_{IMMHC}	440 (440)	53.6 ± 28.4	155 (35.6%)	–	440 (100%)	–	–	–
BCG_{SSPH}	794 (794)	41.3 ± 28.0	279 (69.2%)	–	794 (100%)	–	–	–
BCG (Total)	2179 (2179)	44.7 ± 31.4	1127 (63.2%)	–	2179 (100%)	–	–	–

Continuous variables are presented as mean ± 2 SD, while categorical variables are summarized as counts and corresponding percentages, computed at the recording level after excluding recordings

with unknown classes. “Others” includes individuals identifying as multiracial, Pacific Islander, American Indian, or other racial/ethnic groups. American Indian 和 others. ABC, Apnea, Bariatric Surgery, and CPAP Study; ApoE, Sleep Disordered Breathing: ApoE and Lipid Metabolism; APPLES, Apnea Positive Pressure Long-term Efficacy Study; BestAIR, Best Apnea Interventions in Research; CCSHS, Cleveland Children’s Sleep and Health Study; CFS, Cleveland Family Study; CHAT, Childhood Adenotonsillectomy Trial; HeartBEAT, Heart Biomarker Evaluation in Apnea Treatment; HomePAP, Home Positive Airway Pressure; HSP, Human Sleep Project; MESA, Multi-Ethnic Study of Atherosclerosis; MNC, Mignot Nature Communications; MrOS, MrOS Sleep Study; MSP, Maternal Sleep in Pregnancy and the Fetus; NCHSDB, NCH Sleep DataBank; nuMoM2b, Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be; PATS, Pediatric Adenotonsillectomy Trial of Snoring; SOF, the Study of Osteoporotic Fractures; STAGES, Stanford Technology Analytics and Genomics in Sleep; AHGUCM, the Affiliated Hospital of Gansu University of Chinese Medicine; BHH, Beijing Huilongguan Hospital; IMMHC, Inner Mongolia Mental Health Center; SHSMU, Shenzhen Hospital of Southern Medical University; BTH, Beijing Tongren Hospital, Capital Medical University; SHHS, the Sleep Heart Health Study; WSC, the Wisconsin Sleep Cohort. RH, the Ruijin Hospital; SSPH, Shanghai Sixth People’s Hospital; SRRSH, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine; RH, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine; SAHSU, the second Affiliated Hospital of Soochow University; ABHGMU, the Affiliated Brain Hospital of Guangzhou Medical University; CUHKMC, the Chinese University of Hong Kong Medical Center; WCH, the West China Hospital, Sichuan University.

Cardiorespiratory Signals Capture Core Features of Sleep Physiology

SleepFounder was pretrained on physiology-inspired pretext tasks, using cardiorespiratory signals (heartbeat and respiration) to reconstruct EEG spectrograms and predict oxygen desaturation. Qualitatively, the reconstructed outputs closely resembled their target physiological patterns. As shown in representative cases (**Figure 2**a,b), the predicted oxygen desaturation curves accurately tracked the temporal dynamics of measured events across whole nights of sleep, while the reconstructed EEG spectrograms reproduced both the temporal structure and frequency-band composition of the target recordings.

Quantitatively, evaluation across the internal dataset and two external cohorts (SHHS and WSC) confirmed that these signals contain sufficient information to approximate neural and oxygenation dynamics during sleep (**Figure 2**c). For EEG spectrogram reconstruction, Pearson correlations reached 0.851 (0.847-0.855) internally, 0.824 (0.822-0.827) in SHHS, and 0.910 (0.909-0.912) in WSC. Oxygen desaturation prediction showed consistent associations, with correlations of 0.719 (0.703-0.735), 0.614 (0.604-0.624), and 0.660 (0.644-0.675), respectively. Low MAE values further underscored that cardiorespiratory dynamics capture key aspects of sleep physiology.

Beyond pretraining, we benchmarked SleepFounder against task-specific models and input modalities (e.g., full PSG, ECG-only) across diverse downstream tasks (Supplementary Figure 2, Supplementary Figure 3 and Supplementary Table 3). Despite relying solely on cardiorespiratory signals, SleepFounder achieved performance that was better than, comparable to, or only slightly below that of specialized approaches, demonstrating that fundamental cardiorespiratory activity carries rich physiological information that generalizes across tasks.

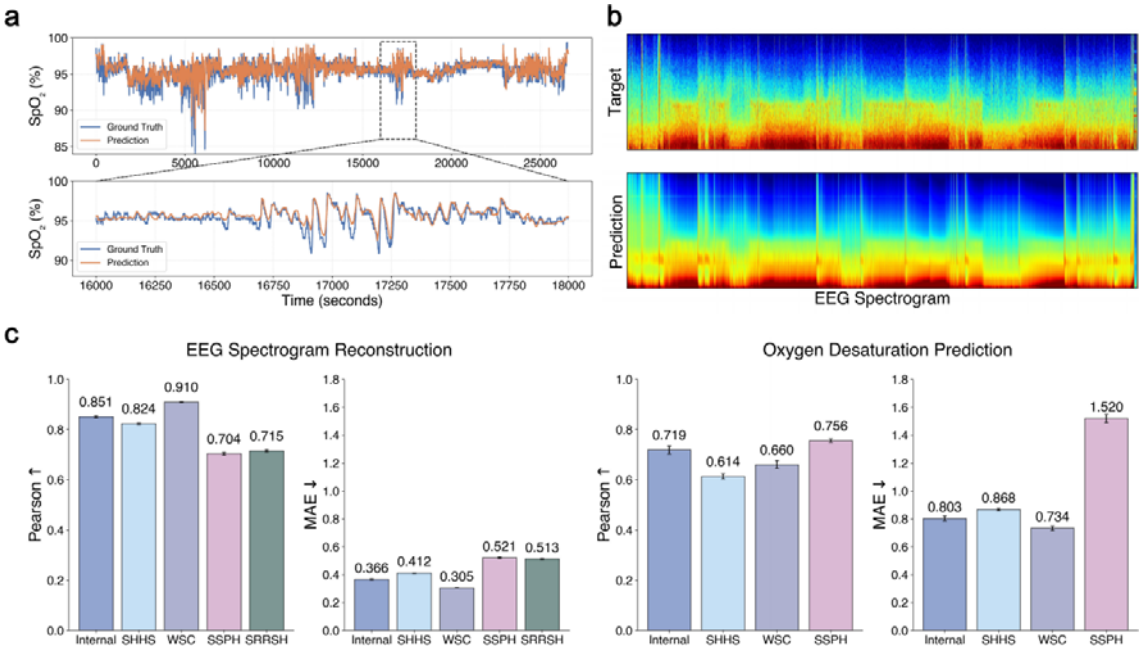


Figure 2. Performance of SleepFounder on physiologically relevant pretext tasks. (a) Example of oxygen desaturation prediction from a single subject's record, showing close alignment between predicted and target

curves. The predicted SpO_2 curve was obtained by subtracting the predicted desaturation values from the local window maximum baseline derived from the ground-truth SpO_2. (b) EEG spectrogram reconstruction from the same record, demonstrating that the model can recover sleep-related neural dynamics using only cardiorespiratory signals. (c) Quantitative evaluation of both pretext tasks on the internal dataset and two external cohorts (SHHS and WSC), showing strong and consistent correlations and low mean absolute errors using only cardiorespiratory signals. EEG, electroencephalography; SHHS, the Sleep Heart Health Study; WSC, the Wisconsin Sleep Cohort.

SleepFounder Demonstrates Superior Performance on Sleep Analysis and Demographic Profiling

We validated SleepFounder on four external datasets, including SHHS, WSC, SRRSH, and SSPH. The former two originated from the United States and the latter two from China, and all were excluded from pretraining. These datasets were used to evaluate both conventional sleep analysis tasks (five-class sleep staging and OSA detection) and emerging sleep health profiling tasks (age prediction and sex classification).

Across all four tasks, SleepFounder demonstrated strong capability and generally achieved state-of-the-art performance (**Figure 3**a and Supplementary Figure 1), obtaining the best results in 14 of 17 dataset-task pairs. In internal validation, it achieved a Cohen's kappa of 0.693 (95% CI: 0.687-0.698) for five-class sleep staging and an AUROC of 0.950 (0.943-0.956) for OSA detection, as well as an MAE of 4.890 years (4.810-4.960) for age prediction and an AUROC of 0.943 (0.937-0.950) for sex classification. Each metric represented the best performance among all compared models. In external validation across four independent cohorts, SleepFounder outperformed baseline models in 10 of 13 dataset-task pairs and achieved the best average performance across all four tasks (**Figure 3**a), with an average Cohen's kappa of 0.671 (0.668-0.673) for sleep staging, an AUROC of 0.917 (0.912-0.922) for OSA detection, an MAE of 6.727 (6.684-6.771) for age prediction, and an AUROC of 0.865 (0.860-0.870) for sex classification, underscoring its strong generalizability across populations. Furthermore, compared with both end-to-end counterparts and SleepFM-CR, SleepFounder achieved superior performance in 13 of 13 and 10 of 13 external evaluations, respectively, demonstrating the effectiveness and robustness of its physiology-inspired pretraining strategy. Collectively, these results establish SleepFounder as a reliable and generalizable foundation model for both conventional and emerging applications in sleep health.

We further evaluated the data efficiency of SleepFounder. The four external datasets were split in an 8:1:1 ratio for tuning, validation, and held-out testing. A subset of the tuning data was used to fine-tune the model and examine the effect of training sample size on performance. Remarkably, the model maintained strong performance even when trained with only a fraction of labeled data, often matching the results of competing models with as little as 5-50% of the labels, averaging only 22% (**Figure 3**b). This data efficiency was most evident in sleep staging and OSA detection, where comparable accuracy was achieved using merely 5-16% of the data. Notably, even under extremely low-resource conditions, SleepFounder delivered clinically meaningful results. For example, when trained with only 20 labeled samples, it achieved Cohen's kappa values of 0.663 and 0.595 in SHHS and WSC for sleep staging, and AUROC values of 0.900 and 0.944 in SHHS and SSPH for OSA detection. These findings demonstrate that the representations learned during pretraining are highly transferable, physiologically grounded, and enable effective adaptation to new datasets and clinical environments even under severe data

limitations.

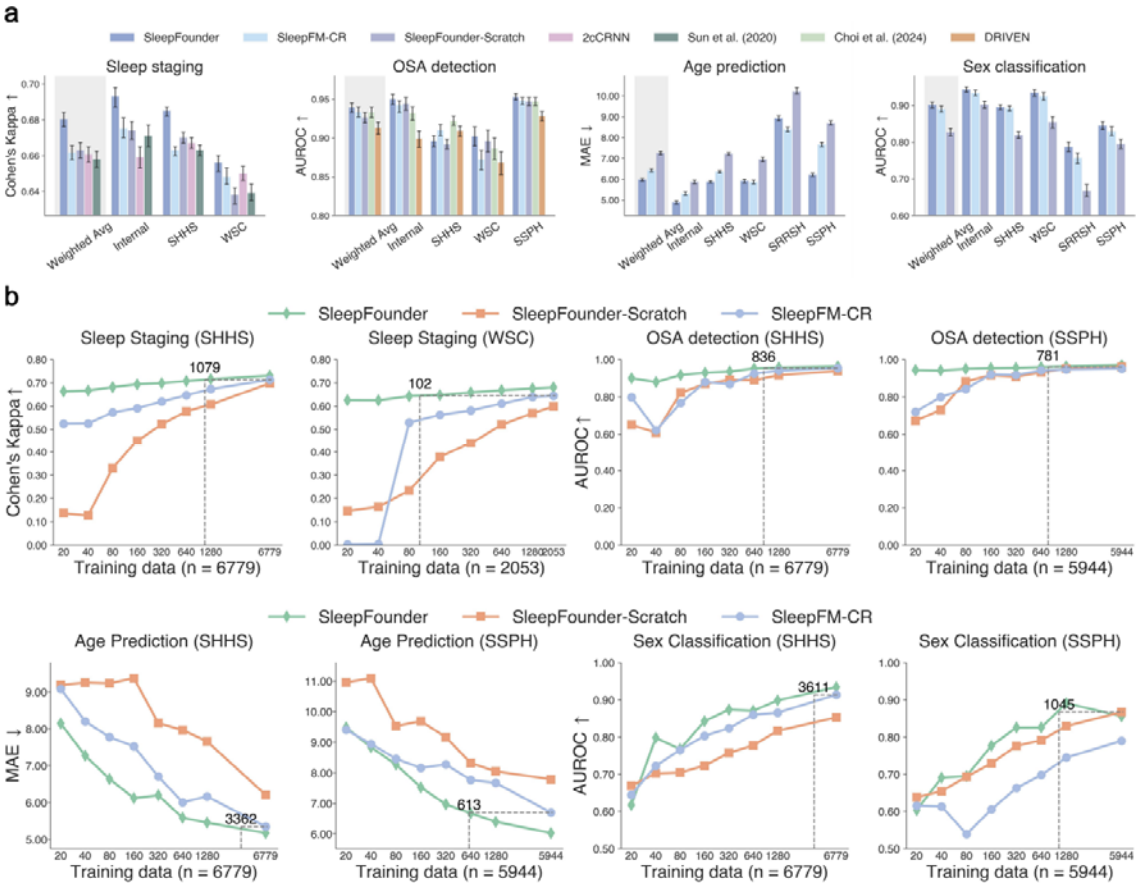


Figure 3. Performance of SleepFounder compared with baseline models across downstream tasks. (a) Internal and external validation results for sleep analysis and demographic profiling tasks, showing that SleepFounder achieved performance comparable to or exceeding that of baseline models in both domains. SleepFM-CR represents a modified version of SleepFM that was adapted to match our input modalities and evaluation protocol, thereby enabling a fair and meaningful comparison with SleepFounder. SleepFounder-Scratch refers to a randomly initialized model with the same architecture as SleepFounder, serving as a control to isolate the contribution of large-scale pretraining. (b) Under limited data conditions, SleepFounder exhibited markedly higher data efficiency, achieving both faster convergence and higher overall performance than baseline models across all evaluated tasks. AUROC, area under the receiver operating characteristic curve; MAE, mean absolute error; SHHS, the Sleep Heart Health Study; WSC, the Wisconsin Sleep Cohort; SRRSH, Zhejiang Sir Run Run Shaw Hospital; SSPH, Shanghai Sixth People's Hospital.

SleepFounder Demonstrates Clinical Utility in Multi-Organ Disease Detection

Sleep is a fundamental physiological state in which multiple organs are dynamically regulated and interact. Building on this premise, we evaluated the ability of SleepFounder to extend beyond conventional sleep characterization toward comprehensive multi-organ health assessment using cardiorespiratory signals collected during sleep. For evaluation, SleepFounder was fine-tuned on the pretraining dataset with diagnostic labels covering multiple physiological systems and disease categories. It was then assessed on three levels of generalization: a held-out internal test set, a

temporally external HSP cohort, and four independent external cohorts (SHHS, WSC, SSPH, and SRRSH), all excluded from both pretraining and fine-tuning.

As shown in **Figure 4**, across eight physiological systems and seventeen disease categories, SleepFounder achieved clinically meaningful detection performance, with detailed results provided in Supplementary Table 1. On the held-out internal test set, AUROCs were 0.943 (0.917-0.966) for PD in the neurological system, 0.886 (0.841-0.928) for GERD in the digestive system; 0.881 (0.831-0.922) for heart failure, 0.814 (0.786-0.841) for CHD, 0.776 (0.735-0.818) for cerebrovascular disease, 0.768 (0.683-0.849) for atrial fibrillation and flutter (AF), and 0.721 (0.698-0.745) for hypertension in the cardiovascular system; 0.830 (0.784-0.874) for high cholesterol and 0.785 (0.755-0.811) for diabetes in the endocrine and metabolic system; 0.810 (0.755-0.861) for chronic pain and 0.736 for arthritis (0.695-0.777) in the musculoskeletal system; 0.744 (0.677-0.803) for chronic obstructive pulmonary disease (COPD) in the respiratory system; 0.811 (0.736-0.878) for bipolar disorder, 0.727 (0.694-0.760) for depression, 0.724 (0.656-0.778) for post-traumatic stress disorder (PTSD), and 0.717 (0.680-0.753) for anxiety in mental and behavioral disorders; and 0.737 (0.656-0.813) for urologic or kidney disease in the genitourinary system. These results highlight broad and clinically relevant coverage across major organ systems.

Model performance generalized well to external cohorts. In the temporally external HSP cohort, disease detection performance remained robust across multiple systems, with AUROCs ranging from 0.688 to 0.851 across cardiovascular, metabolic, musculoskeletal, respiratory, mental, and genitourinary conditions. Across independent cohorts from both the United States and China, SleepFounder consistently retained discriminative ability. For cardiometabolic disorders, heart failure was predicted with AUROCs of 0.815 (0.766-0.856) in SHHS and 0.803 (0.689-0.902) in WSC; diabetes with AUROCs of 0.722-0.765 across SHHS, WSC, SSPH, and SRRSH; CHD with 0.718-0.760 across SHHS, WSC, and SRRSH; and hypertension with 0.672-0.733 across SHHS, WSC, SSPH, and SRRSH. Beyond cardiometabolic conditions, SleepFounder also demonstrated robustness in neurological and respiratory outcomes, achieving an AUROC of 0.816 (0.646, 0.930) for PD in SHHS and 0.722-0.740 for COPD across SHHS and WSC.

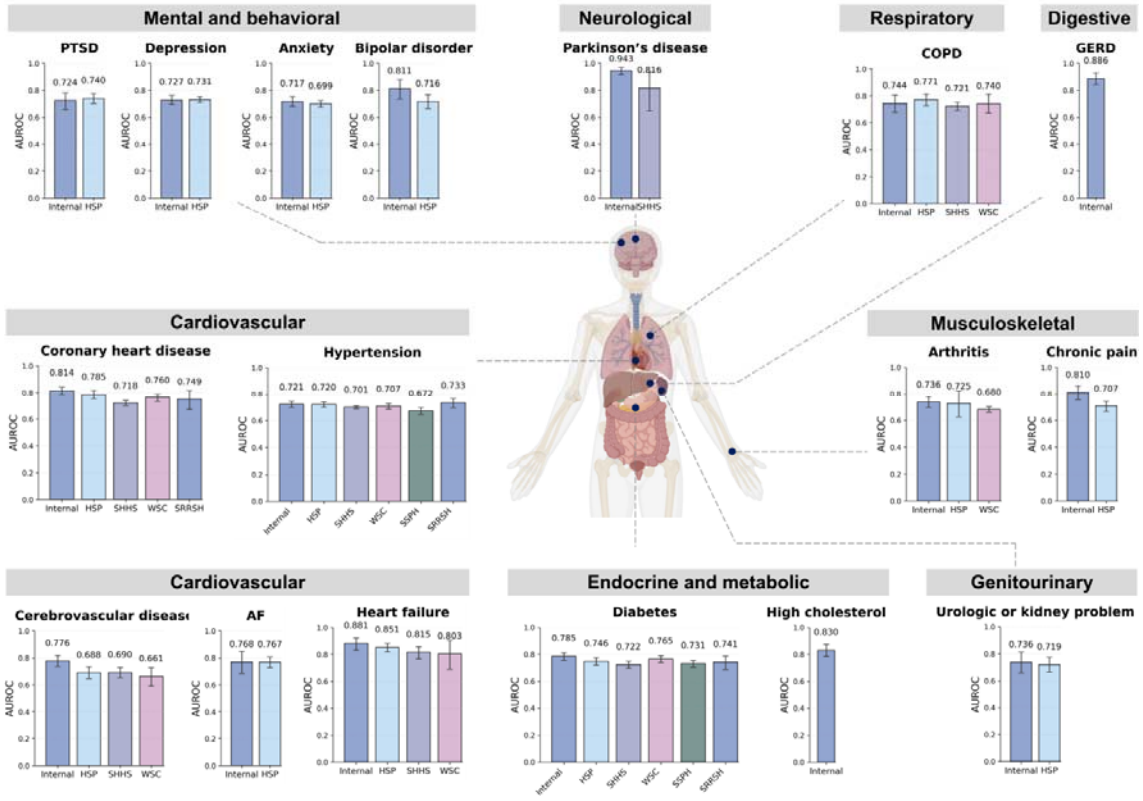


Figure 4. Performance of SleepFounder in detecting multi-organ diseases across internal and external cohorts. Bars show AUROC for each disease grouped by organ system. PTSD, post-traumatic stress disorder; COPD, chronic obstructive pulmonary disease; GERD, gastroesophageal reflux disease; AF, atrial fibrillation/flutter.

SleepFounder Provides Prognostic Value in Future Disease Prediction

Beyond baseline disease detection, we further evaluated SleepFounder's ability to predict future disease outcomes. In the SHHS cohort, which was excluded from pretraining, outcome labels were available for six conditions: angina, CVD death, congestive heart failure, CHD death, myocardial infarction, and stroke. Across these endpoints, prognostic performance was consistent, with 6-year AUROCs ranging from 0.700 to 0.875 and C-indices from 0.673 to 0.888 (**Figure 5**a). The highest discriminative ability was observed for mortality endpoints, including CVD death (AUROC 0.875 [0.831-0.911]; C-index 0.837 [0.806-0.865]) and CHD death (AUROC 0.865 [0.815-0.912]; C-index 0.888 [0.797-0.873]). Stroke (AUROC 0.775 [0.714-0.836]; C-index 0.740 [0.686-0.788]) and angina (AUROC 0.732 [0.687-0.786]; C-index 0.734 [0.684-0.780]) also demonstrated robust performance, while congestive heart failure and myocardial infarction showed moderate but reliable discrimination (AUROCs 0.700-0.730; C-indices 0.673-0.717). Kaplan-Meier survival analysis further confirmed SleepFounder's prognostic capability (**Figure 5**b). Participants stratified into the high-risk group exhibited significantly higher event incidence across all six outcomes (log-rank $P < 0.001$), with hazard ratios ranging from 3.73 to 15.69.

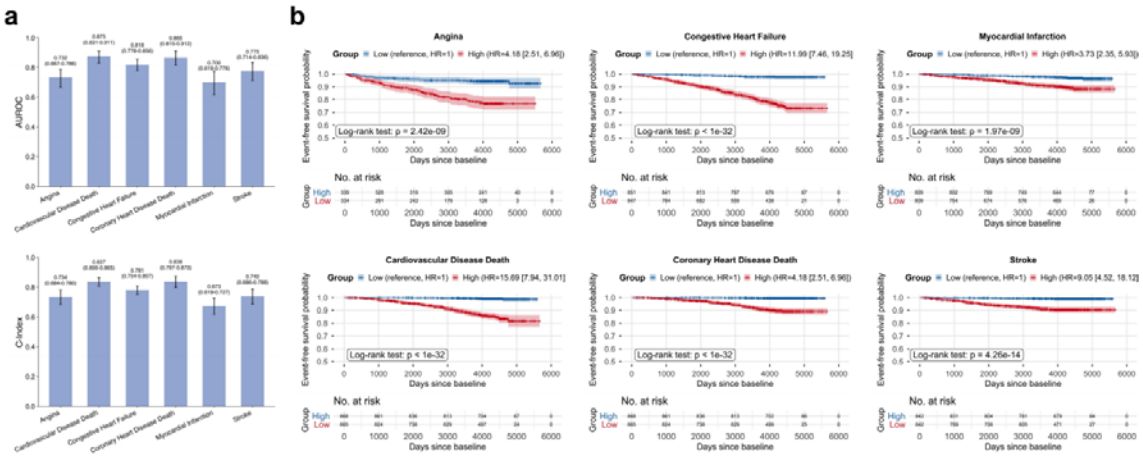


Figure 5. Predictive performance of SleepFounder for future disease prediction in the SHHS cohort (excluded from pretraining). (a) Model performance across six disease outcomes including angina, CVD death, congestive heart failure, coronary heart disease death, myocardial infarction, and stroke, evaluated using 6-year AUROC and C-index, with error bars indicating 95% confidence intervals. (b) Kaplan-Meier survival curves illustrating risk stratification based on predicted risk scores, with participants divided into high- and low-risk groups by the median predicted risk. High-risk participants showed significantly higher event incidence across all outcomes (log-rank $P < 0.001$). AUROC, area under the receiver operating characteristic curve; C-index, concordance index; HR, hazard ratio.

SleepFounder Achieves Robust Performance on Real-world BCG Study

SleepFounder was pretrained on cardiorespiratory signals extracted from PSG recordings. To evaluate whether the model itself can be directly applied beyond clinical PSG settings to commonly used home-based devices that capture cardiorespiratory activity, we conducted a real-world study using multi-center BCG signals collected with our previously developed sleep mat system as a representative example.

For cross-modal agreement assessment, we analyzed recordings where PSG and BCG were acquired simultaneously. As illustrated in **Figure 6**a and **Figure 6**b, BCG-derived respiratory and heartbeat signals showed strong concordance with their PSG counterparts, with respiratory rate achieving near-perfect alignment (Pearson correlation coefficient = 0.94, MAE = 0.29 breaths/min) and heart rate also demonstrating high consistency (Pearson correlation coefficient = 0.91, MAE = 1.84 beats/min). Distribution analyses further confirmed robust cross-modal correlations (**Figure 6**c). These results indicate that the cardiorespiratory signals used to train SleepFounder are reliably captured by BCG devices, thereby ensuring direct applicability of the model in home-based monitoring.

To further evaluate the real-world applicability of SleepFounder, we performed external validation using BCG datasets collected from three independent clinical centers in China, all of which were excluded from both pretraining and fine-tuning. Remarkably, without any additional adaptation, SleepFounder demonstrated strong and consistent performance across four downstream tasks: a Cohen's Kappa of 0.617 (0.610-0.623) for five-class sleep staging, an AUROC of 0.952 (0.943-0.960) for OSA detection, an MAE of 7.580 years (7.440-7.730) for age prediction, and an AUROC of 0.802 (0.781-0.820) for sex classification (**Figure 6**d). Moreover, when compared with multiple baseline models, SleepFounder

consistently outperformed competing approaches across all tasks, underscoring its robustness and direct deployability for real-world, home-based sleep monitoring (**Figure 6**e).

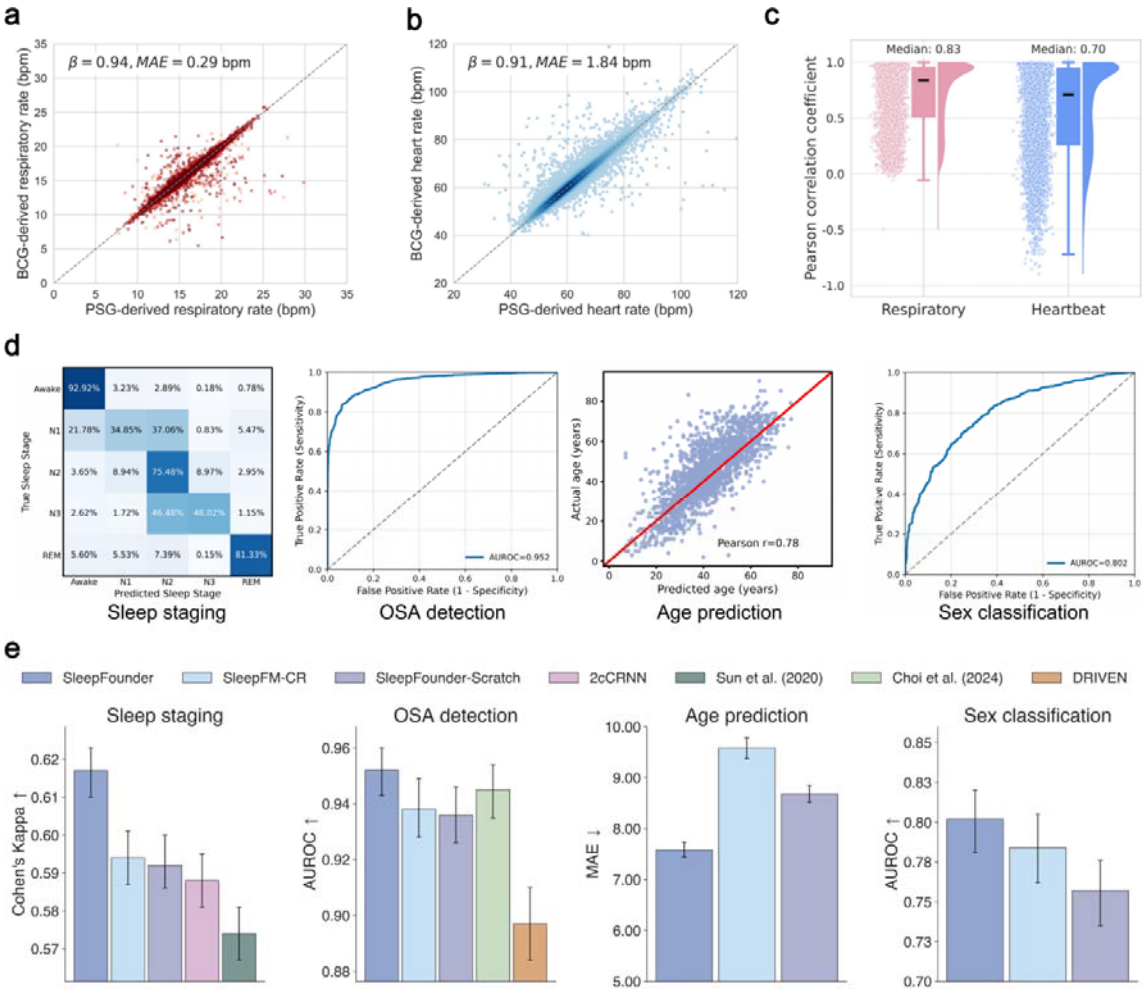


Figure 6. Real-world external validation of SleepFounder on multi-center BCG data. (a-b) Density scatter plots comparing respiratory and heart rates estimated from PSG and BCG signals. (c) Raincloud plots showing the distribution of Pearson correlation coefficients across PSG-BCG pairs for respiratory and heartbeat signals. (d) Performance of SleepFounder across downstream tasks. For sleep staging, a row-normalized confusion matrix illustrates the distribution of predicted sleep stages. The performance on OSA detection and sex classification is summarized using receiver operating characteristic curves, which demonstrate the model's discriminative performance. For age prediction, a scatter plot presents the relationship between predicted and actual ages, along with the Pearson correlation coefficient (r). (e) Performance of SleepFounder versus baseline models across downstream tasks in the external, multi-center BCG dataset. SleepFM-CR represents a modified version of SleepFM that was adapted to match our input modalities and evaluation protocol, thereby enabling a fair and meaningful comparison with SleepFounder. SleepFounder-Scratch denotes a randomly initialized model with the same architecture as SleepFounder, serving as a control to assess the contribution of large-scale pretraining. PSG, Polysomnography; BCG, ballistocardiography; bpm, breaths/beats per minute; β, Pearson correlation coefficient; MAE, mean absolute error; OSA, obstructive sleep apnea; AHI, apnea-hypopnea index; AUROC, area under the receiver operating characteristic curve.

Discussion

In this study, we introduce SleepFounder, a foundation model that leverages widely available cardiorespiratory signals, which are shared physiological modalities in both clinical PSG and home-based systems such as BCG, to enable sleep-centered health assessment. Compared with the previously developed PSG-based sleep foundation model[34], SleepFounder offers greater potential for real-world deployment by relying on practical and easily obtainable input signals.

Overall, SleepFounder consistently achieved strong performance with state-of-the-art performance compared with both self-supervised[34] and task-specific[16,38,39,40] baseline models, often matching or surpassing them while being trained on only 20% of the data, demonstrating strong data efficiency (**Figure 3**). SleepFounder also outperformed previously published approaches. For instance, in sleep staging, it achieved a five-class Cohen's Kappa of 0.685 (0.682-0.687) on the SHHS dataset, outperforming a prior cardiorespiratory-based study[41] that reported a four-class Cohen's Kappa of 0.578 (0.576-0.581). In age prediction, SleepFounder likewise surpassed a previously reported EEG-based model[42], attaining a MAE of 5.866 (5.812-5.920) years on the full SHHS dataset, versus 8.2 years on SHHS1 and 9.4 years on SHHS2. This was superior relative to models based on ECG and respiratory signals[43], which achieved MAEs of 10.4 and 8.1 years, respectively. These improvements can be attributed to the substantially larger dataset and the pretraining strategy employed in SleepFounder.

Given the strong link between sleep and multi-organ health, we evaluated the capacity of SleepFounder in detection of health outcomes across eight physiological systems covering seventeen disease categories (**Figure 4**, Supplementary Table 2). Across these systems, SleepFounder exhibited strong and physiologically coherent performance. The cardiovascular domain showed the most pronounced results, with AUROCs exceeding 0.7 for heart failure, CHD, and AF across all external datasets, underscoring its clinical relevance for CVD, which remains a major global health burden[44]. In the neurological domain, SleepFounder achieved its highest discriminative performance for PD, with an AUROC of 0.943 (0.917-0.966) on the internal test set and 0.816 (0.646-0.930) on the external SHHS dataset. A prior supervised learning approach achieved an AUROC of 0.857 for PD detection using breathing signals from the SHHS dataset[45]. Although SleepFounder exhibited a slightly lower AUROC, its PD-positive training samples were exclusively derived from the Chinese population, and the number of positive cases was approximately one third of that used in the previous study. Overall, the comparable external performance underscores SleepFounder's strong cross-ethnic generalizability despite being a broad, multi-disease model and highlights the efficiency of our pretrain–finetune framework, which substantially reduces the number of labeled cases required. Beyond these expected associations, SleepFounder also revealed previously underrecognized yet physiologically interpretable findings. For example, in the digestive system, the model achieved an AUROC of 0.886 (0.841-0.928) for GERD, reflecting the influence of nocturnal reflux events that disturb sleep and induce distinctive cardiorespiratory alterations such as irregular breathing, reduced heart rate variability, and elevated autonomic arousal[46,47], which were effectively captured by the model.

In addition, SleepFounder achieved strong prognostic performance for future disease prediction within the SHHS cohort (**Figure 5**). Compared with SleepFM[34], which was trained on PSG data with demographic information (age and sex), SleepFounder achieved comparable performance across future

cardiovascular outcomes using only cardiorespiratory signals, with an average C-index of 0.767 versus 0.807. When age and sex information were incorporated, SleepFounder further improved to an average C-index of 0.816, surpassing SleepFM (Supplementary Figure 3). For context, the SCORE2 clinical risk prediction algorithm[48] achieved a C-index of 0.739 in its derivation cohort, whereas the Framingham Risk Score (Wilson) reported a median C-index of 0.71 for CVD prediction in a meta-analysis[49]. Although these results are not directly comparable, SleepFounder achieved a comparable level of performance using only one night of passively acquired cardiorespiratory signals, in contrast to conventional scores that depend on multiple clinical and biochemical variables (e.g., high-density lipoprotein cholesterol) typically collected in healthcare settings.

Beyond these findings, external validation on multi-center BCG data collected from three independent institutions in China demonstrated the strong real-world applicability of SleepFounder for zero-burden home monitoring (**Figure 6**). The model preserved nearly 94% of its downstream performance compared with PSG-based external validation, indicating consistent stability across device modalities. Taken together, these findings underscore the feasibility of deploying large-scale, cardiorespiratory-based foundation models for practical, low-cost, and contact-free sleep monitoring in real-world settings.

Despite these promising results, several limitations warrant consideration. First, although SleepFounder underwent extensive validation across multiple large-scale PSG cohorts from both the United States and China, as well as on real-world BCG recordings collected from three independent centers using our custom-developed sleep mat system, pretraining relied exclusively on PSG data. While these validations already demonstrate strong cross-cohort generalizability, further evaluation across broader populations and more diverse environments will be necessary to confirm the model's real-world applicability. Second, although the model exhibited strong performance in disease detection across multiple physiological systems, disease labels were primarily self-reported rather than clinically adjudicated, which may introduce labeling bias. Future efforts incorporating harmonized electronic health records and clinically confirmed outcomes will be crucial to strengthen detection fidelity and translational readiness. Finally, longitudinal cardiovascular prediction was limited to the SHHS cohort; expanding to cohorts with broader diagnostic coverage and extended follow-up will be important to fully characterize prognostic utility. Ongoing large-scale deployment of the BCG-based sleep mat will enable continuous, real-world data acquisition to further evaluate its applicability in home-based preventive care.

In conclusion, SleepFounder demonstrates that cardiorespiratory-based foundation models can enable accurate, scalable, and zero-burden sleep monitoring. Built upon large-scale, multi-ethnic PSG pretraining and validated across both clinical PSG and BCG datasets, the model achieved robust and generalizable performance across a spectrum of sleep and multi-organ health tasks. By integrating accessible cardiorespiratory signals with physiology-informed pretraining, SleepFounder establishes a scalable framework for continuous, real-world, and individualized AI-driven sleep and health assessment. These findings extend foundation-model applications beyond traditional PSG and introduce a new paradigm for precision preventive medicine, in which passive physiological monitoring enables early risk detection, targeted intervention, and proactive health management in out-of-clinic settings.

Acknowledgments

This study was conducted as part of the AISleep 365 Consortium, a multi-center harmonization initiative uniting hospital-based PSG data across China to enable AI-driven and home-based sleep health monitoring. We thank the participating hospitals and investigators of the AISleep 365 Consortium for data sharing and collaboration. This work is supported by the Ministry of Science and Technology of China STI2030-Major Projects (No.2021ZD0201900, 2021ZD0201902), and the National Natural Science Foundation of China (62102008).

The authors gratefully acknowledge the National Sleep Research Resource (NSRR; sleepdata.org), funded by the National Heart, Lung, and Blood Institute (NHLBI), for providing access to de-identified clinical data that supported the development and evaluation of our work. We also thank the Brain Data Science Platform (BDSP) for access to the Human Sleep Project (HSP) dataset, a growing collection of clinical polysomnography recordings from Massachusetts General Hospital.

Declaration of interests

Westover is a co-founder, serves as a scientific advisor and consultant to, and has a personal equity interest in Beacon Biosignals. Thomas discloses: 1) patent and license/royalties from MyCardio, LLC, for the ECG-spectrogram; 2) patent and license/royalties from DeVilbiss-Drive for an auto-CPAP algorithm; 3) consulting for Jazz Pharmaceuticals, Guidepoint Global and GLG Councils.

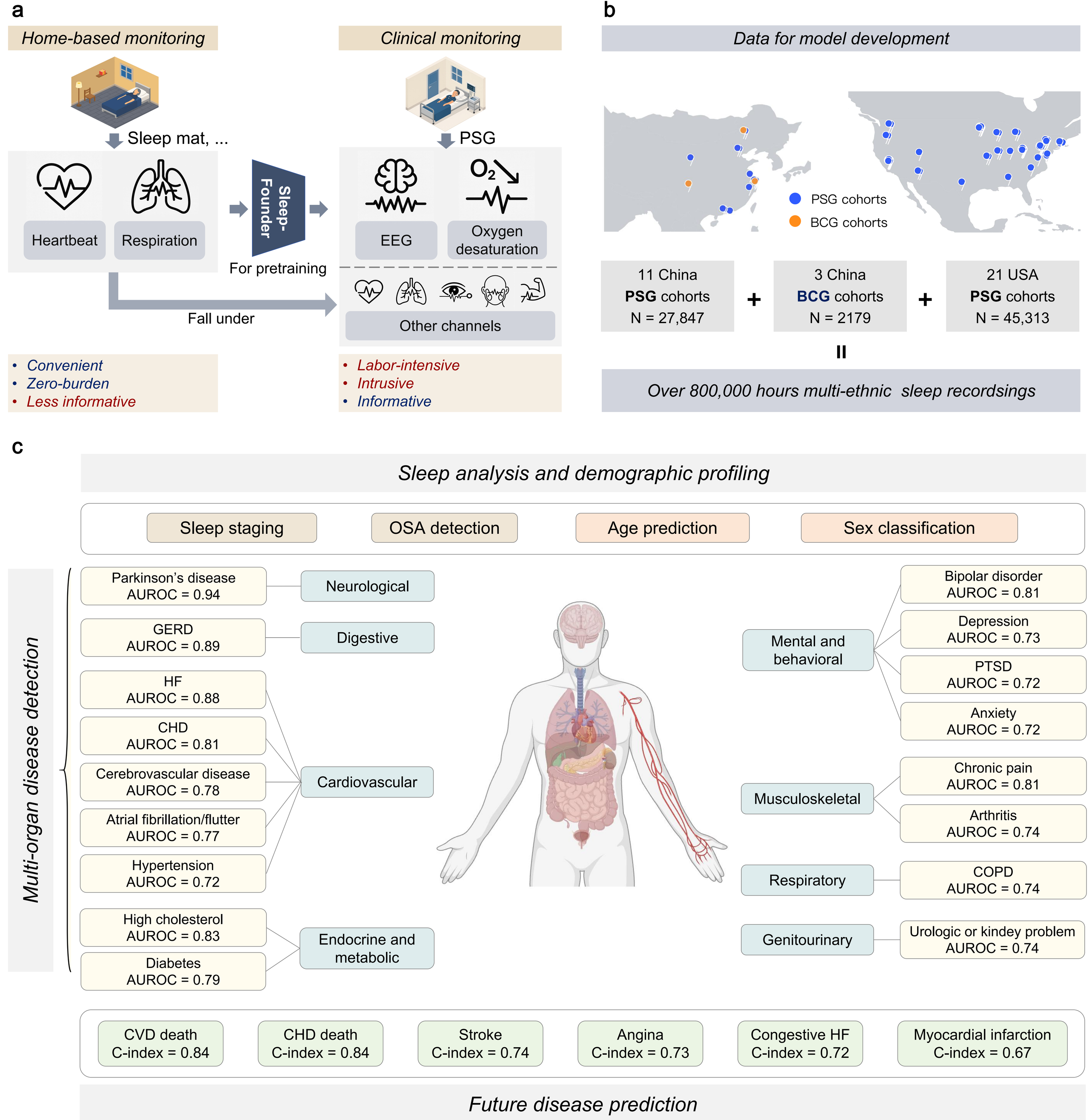
References

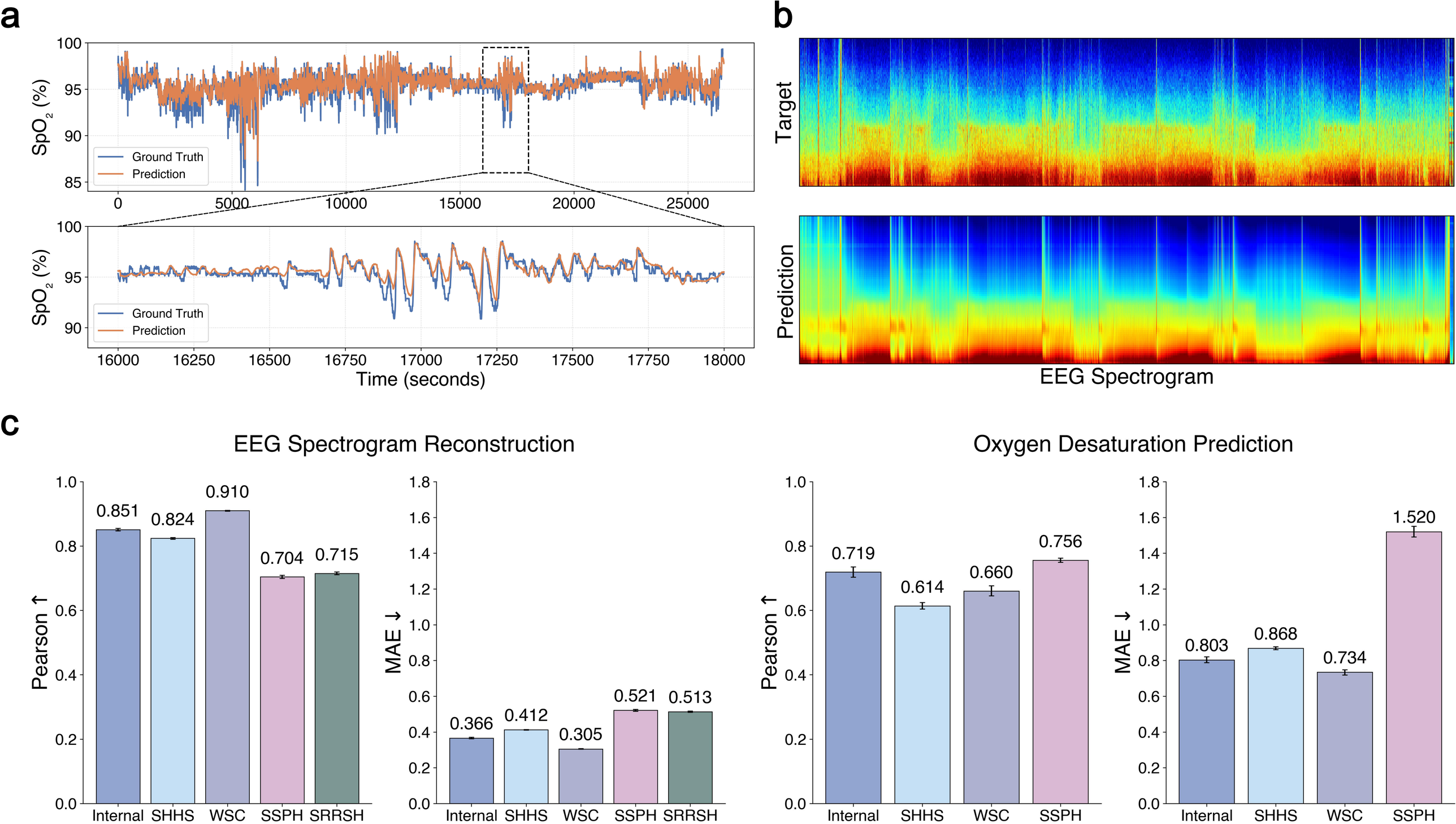
1. Ge L, Guyatt G, Tian J, Pan B, Chang Y, Chen Y, Li H, Zhang J, Li Y, Ling J, Yang K. Insomnia and risk of mortality from all-cause, cardiovascular disease, and cancer: Systematic review and meta-analysis of prospective cohort studies. Sleep Medicine Reviews. 2019 Dec 1;48:101215.
2. Ai S, Zhang J, Zhao G, Wang N, Li G, So HC, Liu Y, Chau SW, Chen J, Tan X, Jia F. Causal associations of short and long sleep durations with 12 cardiovascular diseases: linear and nonlinear Mendelian randomization analyses in UK Biobank. European Heart Journal. 2021 Sep 7;42(34):3349-57.
3. Nikbakhtian S, Reed AB, Obika BD, Morelli D, Cunningham AC, Aral M, Plans D. Accelerometer-derived sleep onset timing and cardiovascular disease incidence: a UK Biobank cohort study. European Heart Journal-Digital Health. 2021 Dec 1;2(4):658-66.
4. Yeghiazarians Y, Jneid H, Tietjens JR, Redline S, Brown DL, El-Sherif N, Mehra R, Bozkurt B, Ndumele CE, Somers VK. Obstructive sleep apnea and cardiovascular disease: a scientific statement from the American Heart Association. Circulation. 2021 Jul 20;144(3):e56-67.
5. Schmid SM, Hallschmid M, Schultes B. The metabolic burden of sleep loss. The lancet Diabetes & endocrinology. 2015 Jan 1;3(1):52-62.
6. Kianersi S, Wang H, Sofer T, Noordam R, Phillips A, Rutter MK, Redline S, Huang T. Association between accelerometer-measured irregular sleep duration and type 2 diabetes risk: a prospective cohort study in the UK biobank. Diabetes care. 2024 Sep 1;47(9):1647-55.
7. Leng Y, Musiek ES, Hu K, Cappuccio FP, Yaffe K. Association between circadian rhythms and neurodegenerative diseases. The Lancet Neurology. 2019 Mar 1;18(3):307-18.
8. Leng Y, McEvoy CT, Allen IE, Yaffe K. Association of sleep-disordered breathing with cognitive function and risk of cognitive impairment: a systematic review and meta-analysis. JAMA neurology. 2017 Oct 1;74(10):1237-45.
9. Palagini L, Hertenstein E, Riemann D, Nissen C. Sleep, insomnia and mental health. Journal of sleep research. 2022 Aug;31(4):e13628.
10. Kim JY, Ko I, Kim DK. Association of obstructive sleep apnea with the risk of affective disorders. JAMA Otolaryngology-Head & Neck Surgery. 2019 Nov 1;145(11):1020-6.
11. Benjafield AV, Ayas NT, Eastwood PR, Heinzer R, Ip MS, Morrell MJ, Nunez CM, Patel SR, Penzel T, Pépin JL, Peppard PE. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. The Lancet respiratory medicine. 2019 Aug 1;7(8):687-98.
12. Watson N, Yu K, Campbell D, Bloudek L, Yee J, Cronin J, Azofra AS, Boskovic N. 0637 Prevalence and Unmet Need of Obstructive Sleep Apnea in the United States. Sleep. 2025 May;48(Supplement_1):A278-.
13. Faria A, Allen AH, Fox N, Ayas N, Laher I. The public health burden of obstructive sleep apnea. Sleep Science. 2021 Sep;14(03):257-65.
14. Li X, Li M, Tian M, Liu Q, Zhou X, Liu H, Li R, Li Z, Dong H, Jia L, Liu Y. Exploring the potential of a new wearable sleep monitoring device for clinical application. Biomedical Signal Processing and Control. 2025 Jan 1;99:106856.
15. Kwon S, Kim H, Yeo WH. Recent advances in wearable sensors and portable electronics for sleep monitoring. iScience. 2021 May 21;24(5).
16. Retamales G, Gavidia ME, Bausch B, Montanari AN, Husch A, Goncalves J. Towards automatic home-based sleep apnea estimation using deep learning. npj Digital Medicine. 2024 Jun 1;7(1):144.
17. Hu S, Liu J, Wang Y, Fu C, Zhu J, Yu H, Yang C. Transparent artificial intelligence-enabled

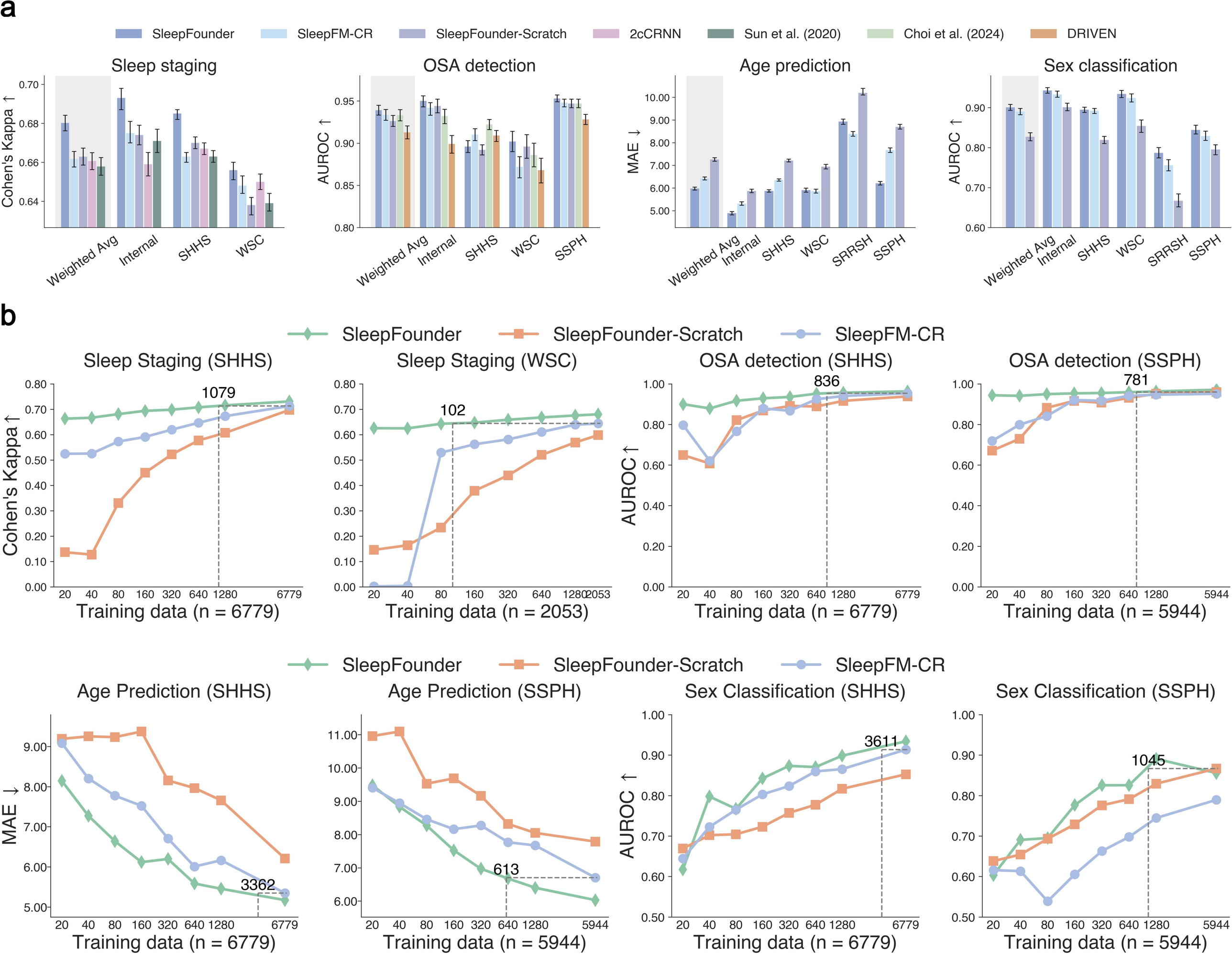
interpretable and interactive sleep apnea assessment across flexible monitoring scenarios. Nature Communications. 2025 Aug 14;16(1):7548.
18. Levy J, Álvarez D, Del Campo F, Behar JA. Deep learning for obstructive sleep apnea diagnosis based on single channel oximetry. Nature Communications. 2023 Aug 12;14(1):4881.
19. Du Y, Gu J, Duan S, Trueb J, Tzavelis A, Shin HS, Arafa H, Li X, Huang Y, Carr AN, Davies CR. A skin-interfaced wireless wearable device and data analytics approach for sleep-stage and disorder detection. Proceedings of the National Academy of Sciences. 2025 Jun 10;122(23):e2501220122.
20. Bakker JP, Ross M, Vasko R, Cerny A, Fonseca P, Jasko J, Shaw E, White DP, Anderer P. Estimating sleep stages using cardiorespiratory signals: validation of a novel algorithm across a wide range of sleep-disordered breathing severity. Journal of Clinical Sleep Medicine. 2021 Jul 1;17(7):1343-54.
21. Anderer P, Ross M, Cerny A, Radha M, Fonseca P. 0436 Deep Learning for Scoring Sleep Based on Signals Available in Home Sleep Apnea Test Studies: Cardiorespiratory Sleep Staging. Sleep. 2020 Apr 1;43:A167.
22. Cerina L, Overeem S, Papini GB, van Dijk JP, Vullings R, van Meulen F, Ross M, Cerny A, Anderer P, Fonseca P. A sleep stage estimation algorithm based on cardiorespiratory signals derived from a suprasternal pressure sensor. Journal of Sleep Research. 2024 Apr;33(2):e14015.
23. Ross M, Fonseca P, Overeem S, Vasko R, Cerny A, Shaw E, Anderer P. Autonomic arousal detection and cardio-respiratory sleep staging improve the accuracy of home sleep apnea tests. Frontiers in Physiology. 2023 Aug 24;14:1254679.
24. Fonseca P, Ross M, Cerny A, Anderer P, Schipper F, Grassi A, van Gilst M, Overeem S. Estimating the severity of obstructive sleep apnea using ecg, respiratory effort and neural networks. IEEE Journal of Biomedical and Health Informatics. 2024 Mar 29;28(7):3895-906.
25. Retamales G, Gavidia ME, Bausch B, Montanari AN, Husch A, Goncalves J. Towards automatic home-based sleep apnea estimation using deep learning. npj Digital Medicine. 2024 Jun 1;7(1):144.
26. Hu S, Liu J, Wang Y, Fu C, Zhu J, Yu H, Yang C. Transparent artificial intelligence-enabled interpretable and interactive sleep apnea assessment across flexible monitoring scenarios. Nature Communications. 2025 Aug 14;16(1):7548.
27. Chen RJ, Ding T, Lu MY, Williamson DF, Jaume G, Song AH, Chen B, Zhang A, Shao D, Shaban M, Williams M. Towards a general-purpose foundation model for computational pathology. Nature medicine. 2024 Mar;30(3):850-62.
28. Vorontsov E, Bozkurt A, Casson A, Shaikovski G, Zelechowski M, Severson K, Zimmermann E, Hall J, Tenenholtz N, Fusi N, Yang E. A foundation model for clinical-grade computational pathology and rare cancers detection. Nature medicine. 2024 Oct;30(10):2924-35.
29. Christensen M, Vukadinovic M, Yuan N, Ouyang D. Vision–language foundation model for echocardiogram interpretation. Nature Medicine. 2024 May;30(5):1481-8.
30. Abbaspourazad S, Elachqar O, Miller A, Emrani S, Nallasamy U, Shapiro I. Large-scale Training of Foundation Models for Wearable Biosignals. InThe Twelfth International Conference on Learning Representations.
31. Tian Y, Li Z, Jin Y, Wang M, Wei X, Zhao L, Liu Y, Liu J, Liu C. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. Cell Reports Medicine. 2024 Dec 17;5(12).
32. Li J, Aguirre AD, Junior VM, Jin J, Liu C, Zhong L, Sun C, Clifford G, Brandon Westover M, Hong S. An Electrocardiogram Foundation Model Built on over 10 Million Recordings. NEJM AI. 2025 Jun 26;2(7):AIoa2401033.

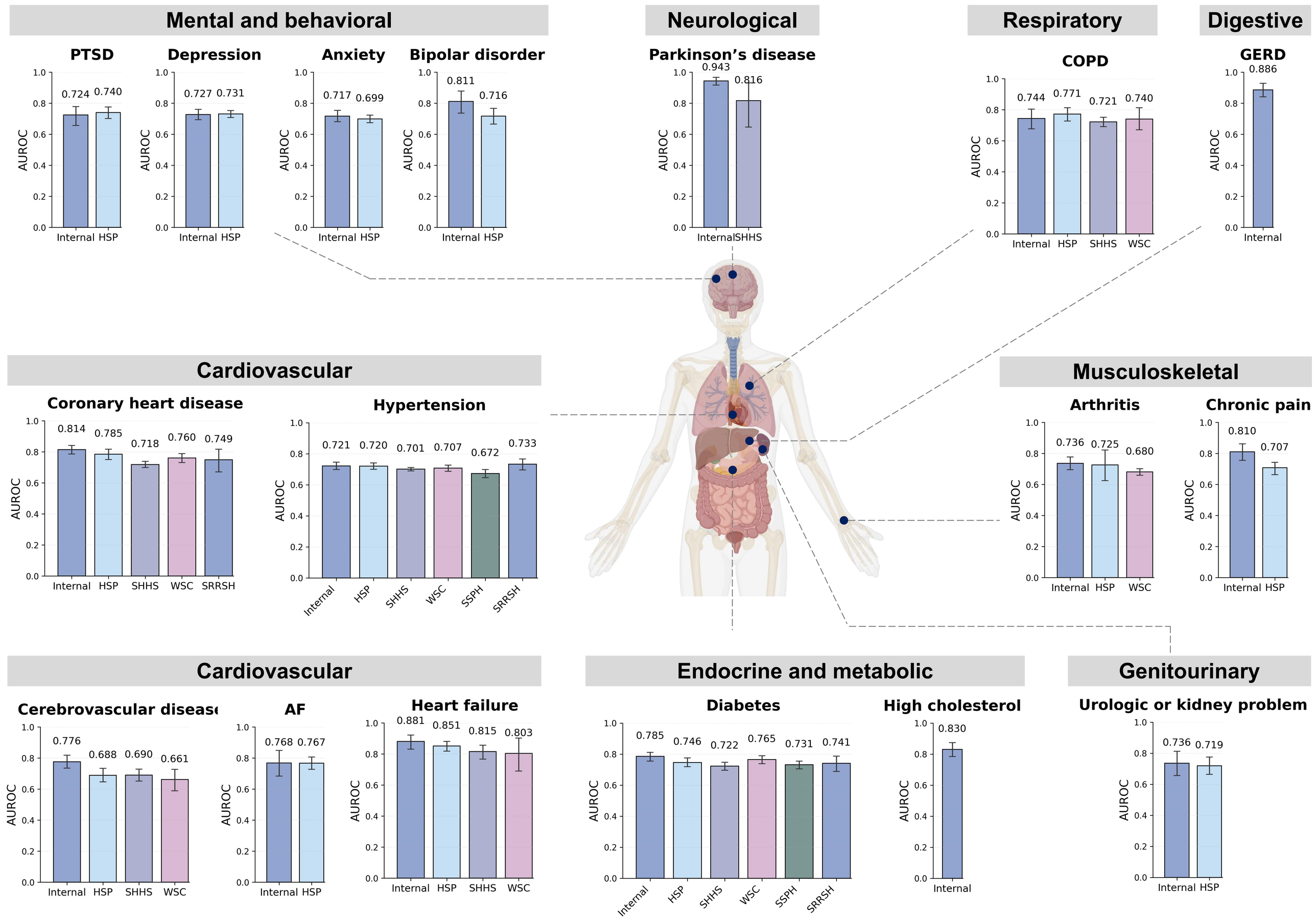
33. Li J, Goldenholz DM, Alkofer M, Sun C, Nascimento FA, Halford JJ, Dean BC, Galanti M, Struck AF, Greenblatt AS, Lam AD. Expert-Level Detection of Epilepsy Markers in EEG on Short and Long Timescales. NEJM AI. 2025 Jun 26;2(7):AIoa2401221.
34. Thapa R, Kjær MR, He B, Covert I, Moore H, Hanif U, Ganjoo G, Westover MB, Jennum P, Brink-Kjær A, Mignot E. A multimodal sleep foundation model developed with 500k hours of sleep recordings for disease predictions. medRxiv. 2025 Feb 9.
35. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).
36. Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing. 2024 Feb 1;568:127063.
37. Chen S, Chen X, Huang W, Lei F, Shan C, Jin Z, Shi Y, Wang Y, Zhao R, Xu X, Lv D. BCGNet: An AI Model Trained on 600K Hours of Sleep Data for a Novel Under-Pillow Contactless Monitoring Device. medRxiv. 2025 Aug 26:2025-08.
38. Goldammer M, Zaunseder S, Brandt MD, Malberg H, Gräßer F. Investigation of automated sleep staging from cardiorespiratory signals regarding clinical applicability and robustness. Biomedical Signal Processing and Control. 2022 Jan 1;71:103047.
39. Sun H, Ganglberger W, Panneerselvam E, Leone MJ, Quadri SA, Goparaju B, Tesh RA, Akeju O, Thomas RJ, Westover MB. Sleep staging from electrocardiography and respiration with deep learning. Sleep. 2020 Jul;43(7):zsz306.
40. Choi JW, Koo DL, Kim DH, Nam H, Lee JH, Hong SN, Kim B. A novel deep learning model for obstructive sleep apnea diagnosis: hybrid CNN-Transformer approach for radar-based detection of apnea-hypopnea events. Sleep. 2024 Dec;47(12):zsae184.
41. Bakker JP, Ross M, Vasko R, Cerny A, Fonseca P, Jasko J, Shaw E, White DP, Anderer P. Estimating sleep stages using cardiorespiratory signals: validation of a novel algorithm across a wide range of sleep-disordered breathing severity. Journal of Clinical Sleep Medicine. 2021 Jul 1;17(7):1343-54.
42. Sun H, Paixao L, Oliva JT, Goparaju B, Carvalho DZ, van Leeuwen KG, Akeju O, Thomas RJ, Cash SS, Bianchi MT, Westover MB. Brain age from the electroencephalogram of sleep. Neurobiology of aging. 2019 Feb 1;74:112-20.
43. Brink-Kjaer A, Leary EB, Sun H, Westover MB, Stone KL, Peppard PE, Lane NE, Cawthon PM, Redline S, Jennum P, Sorensen HB. Age estimation from sleep studies using deep learning predicts life expectancy. NPJ digital medicine. 2022 Jul 22;5(1):103.
44. Vaduganathan M, Mensah GA, Turco JV, Fuster V, Roth GA. The global burden of cardiovascular diseases and risk: a compass for future health. Journal of the American College of Cardiology. 2022 Dec 20;80(25):2361-71.
45. Yang Y, Yuan Y, Zhang G, Wang H, Chen YC, Liu Y, Tarolli CG, Crepeau D, Bukartyk J, Junna MR, Videnovic A. Artificial intelligence-enabled detection and assessment of Parkinson's disease using nocturnal breathing signals. Nature medicine. 2022 Oct;28(10):2207-15.
46. Dickman R, Green C, Fass SS, Quan SF, Dekel R, Risner-Adler S, Fass R. Relationships between sleep quality and pH monitoring findings in persons with gastroesophageal reflux disease. Journal of Clinical Sleep Medicine. 2007 Aug 15;3(5):505-13.
47. Huang Y, Liu J, Xu L, Qi W, Dai J, Wang B, Tian J, Fu X, Yu Y. Exacerbation of symptoms, nocturnal acid reflux, and impaired autonomic function are associated with sleep disturbance in gastroesophageal reflux disease patients. Frontiers in Medicine. 2024 Aug 21;11:1438698.
48. Hageman S, Pennells L, Ojeda F, Kaptoge S, Kuulasmaa K, de Vries T, Xu Z, Kee F, Chung R, Wood

A, Mcevoy JW. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. European Heart Journal. 2021 Jul 1;42(25):2439-54.
49. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiocchia V, Roberts C, Schlüssel MM. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016 May 16;353.

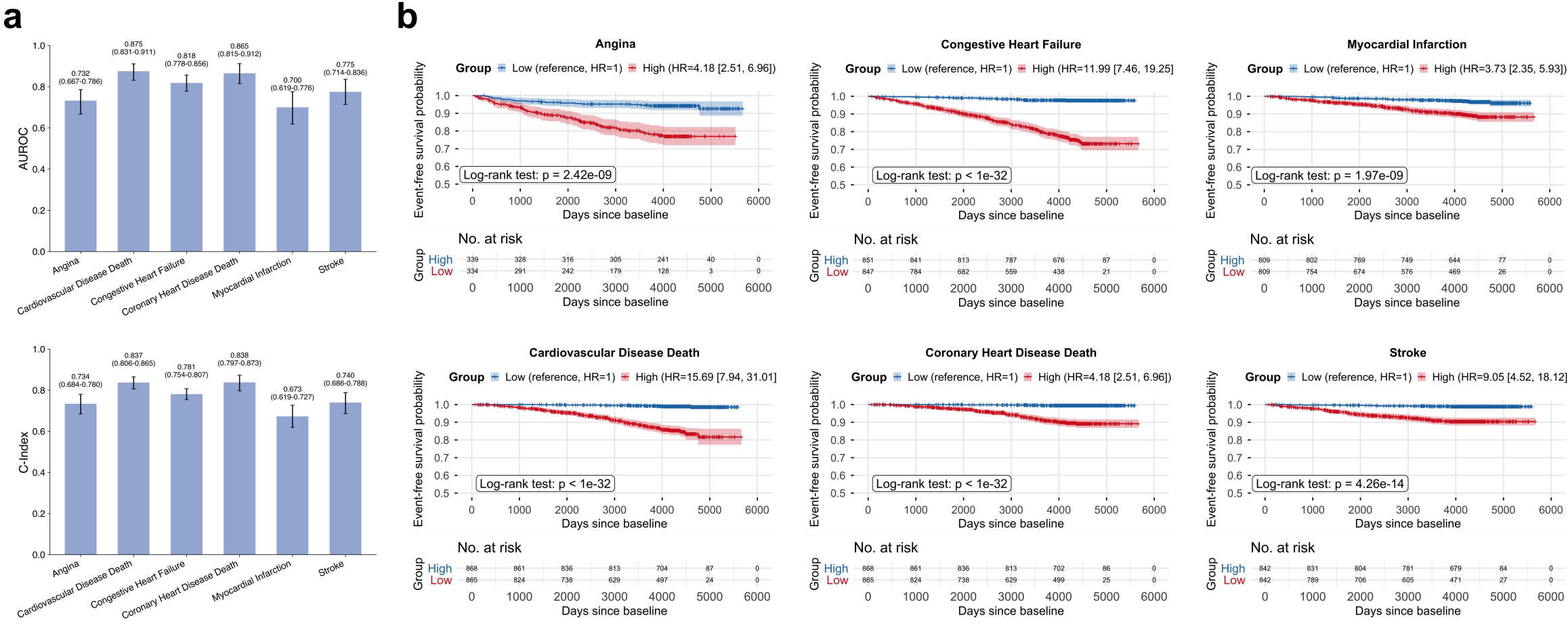

a
Home-based monitoring
Clinical monitoring
Sleep mat, ...
PSG
Heartbeat
Respiration
Sleep-Founder
For pretraining
EEG
Oxygen desaturation
Fall under
Other channels
• Convenient
• Zero-burden
• Less informative
• Labor-intensive
• Intrusive
• Informative
b
Data for model development
PSG cohorts
BCG cohorts
11 China PSG cohorts N = 27,847
+
3 China BCG cohorts N = 2179
+
21 USA PSG cohorts N = 45,313
=
Over 800,000 hours multi-ethnic sleep recordsings
c
Sleep analysis and demographic profiling
Sleep staging
OSA detection
Age prediction
Sex classification
Multi-organ disease detection
Parkinson's disease AUROC = 0.94
Neurological
GERD AUROC = 0.89
Digestive
HF AUROC = 0.88
CHD AUROC = 0.81
Cerebrovascular disease AUROC = 0.78
Atrial fibrillation/flutter AUROC = 0.77
Hypertension AUROC = 0.72
Cardiovascular
High cholesterol AUROC = 0.83
Diabetes AUROC = 0.79
Endocrine and metabolic
Mental and behavioral
Bipolar disorder AUROC = 0.81
Depression AUROC = 0.73
PTSD AUROC = 0.72
Anxiety AUROC = 0.72
Musculoskeletal
Chronic pain AUROC = 0.81
Arthritis AUROC = 0.74
Respiratory
COPD AUROC = 0.74
Genitourinary
Urologic or kindey problem AUROC = 0.74
CVD death C-index = 0.84
CHD death C-index = 0.84
Stroke C-index = 0.74
Angina C-index = 0.73
Congestive HF C-index = 0.72
Myocardial infarction C-index = 0.67
Future disease prediction

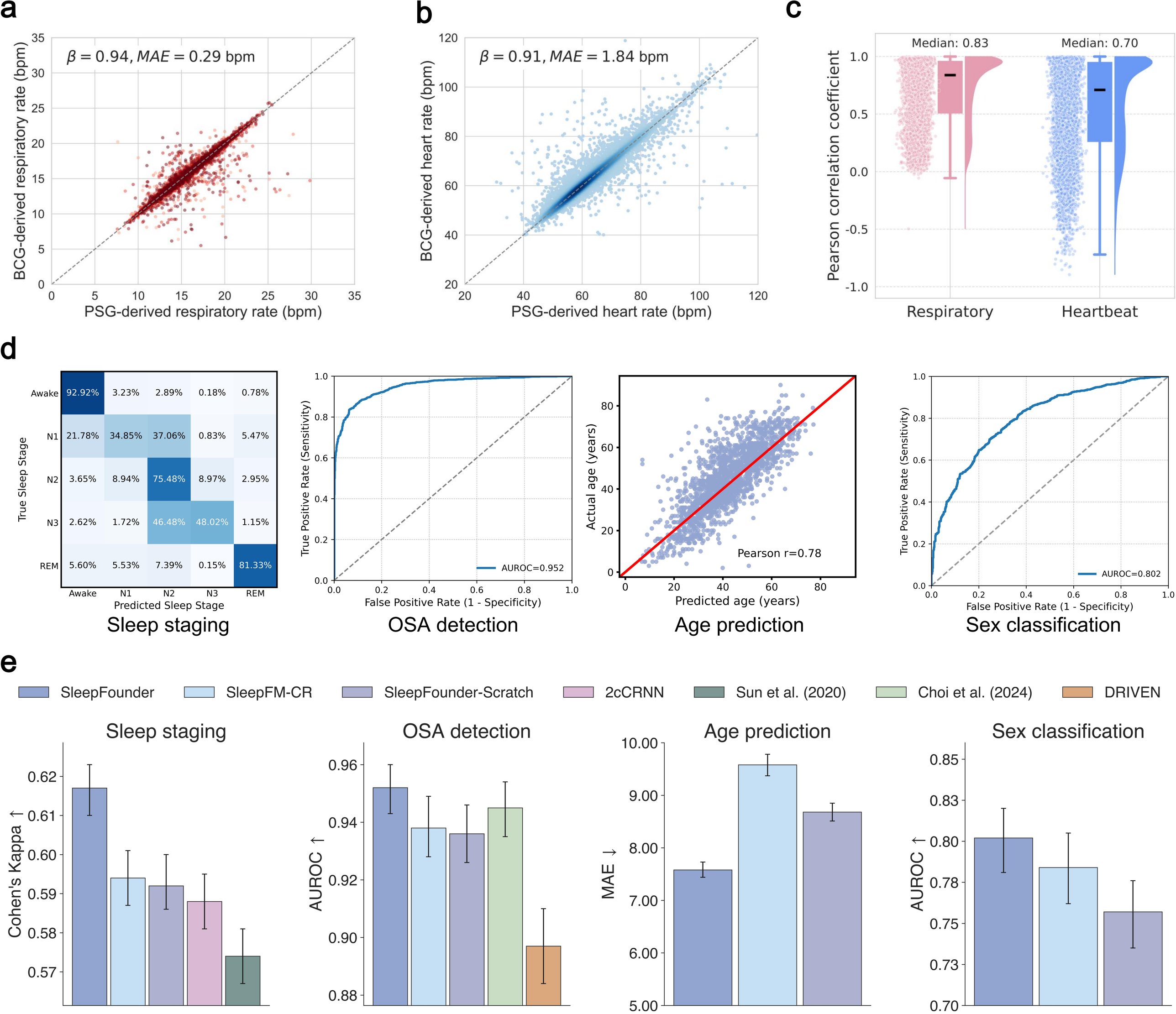

a
SpO2 (%)
Ground Truth
Prediction
Time (seconds)
b
Target
Prediction
EEG Spectrogram
c
EEG Spectrogram Reconstruction
Pearson ↑
MAE ↓
0.851
0.824
0.910
0.704
0.715
0.366
0.412
0.305
0.521
0.513
Internal
SHHS
WSC
SSPH
SRRSH
Oxygen Desaturation Prediction
0.719
0.614
0.660
0.756
0.803
0.868
0.734
1.520


a
SleepFounder
SleepFM-CR
SleepFounder-Scratch
2cCRNN
Sun et al. (2020)
Choi et al. (2024)
DRIVEN
Sleep staging
Cohen's Kappa ↑
OSA detection
AUROC ↑
Age prediction
MAE ↓
Sex classification
AUROC ↑
Weighted Avg
Internal
SHHS
WSC
SSPH
SRRSH
b
SleepFounder
SleepFounder-Scratch
SleepFM-CR
Sleep Staging (SHHS)
1079
Cohen's Kappa ↑
Training data (n = 6779)
Sleep Staging (WSC)
102
Training data (n = 2053)
OSA detection (SHHS)
836
AUROC ↑
Training data (n = 6779)
OSA detection (SSPH)
781
Training data (n = 5944)
Age Prediction (SHHS)
3362
MAE ↓
Training data (n = 6779)
Age Prediction (SSPH)
613
Training data (n = 5944)
Sex Classification (SHHS)
3611
AUROC ↑
Training data (n = 6779)
Sex Classification (SSPH)
1045
Training data (n = 5944)


Mental and behavioral
PTSD
0.724 0.740
Internal HSP
Depression
0.727 0.731
Internal HSP
Anxiety
0.717 0.699
Internal HSP
Bipolar disorder
0.811 0.716
Internal HSP
Neurological
Parkinson's disease
0.943 0.816
Internal SHHS
Respiratory
COPD
0.744 0.771 0.721 0.740
Internal HSP SHHS WSC
Digestive
GERD
0.886
Internal
Cardiovascular
Coronary heart disease
0.814 0.785 0.718 0.760 0.749
Internal HSP SHHS WSC SRRSH
Hypertension
0.721 0.720 0.701 0.707 0.672 0.733
Internal HSP SHHS WSC SSPH SRRSH
Musculoskeletal
Arthritis
0.736 0.725 0.680
Internal HSP WSC
Chronic pain
0.810 0.707
Internal HSP
AUROC

Cardiovascular
Cerebrovascular diseas
0.776 0.688 0.690 0.661
Internal HSP SHHS WSC
AF
0.768 0.767
Internal HSP
Heart failure
0.881 0.851 0.815 0.803
Internal HSP SHHS WSC
Endocrine and metabolic
Diabetes
0.785 0.746 0.722 0.765 0.731 0.741
Internal HSP SHHS WSC SSPH SRRSH
High cholesterol
0.830
Internal
Genitourinary
Urologic or kidney problem
0.736 0.719
Internal HSP
AUROC


a
AUROC
0.732 (0.667-0.786)
0.875 (0.831-0.911)
0.818 (0.778-0.856)
0.865 (0.815-0.912)
0.700 (0.619-0.776)
0.775 (0.714-0.836)
Angina
Cardiovascular Disease Death
Congestive Heart Failure
Coronary Heart Disease Death
Myocardial Infarction
Stroke
C-Index
0.734 (0.684-0.780)
0.837 (0.806-0.865)
0.781 (0.754-0.807)
0.838 (0.797-0.873)
0.673 (0.619-0.727)
0.740 (0.686-0.788)
b
Angina
Group Low (reference, HR=1) High (HR=4.18 [2.51, 6.96])
Log-rank test: p = 2.42e-09
Congestive Heart Failure
Group Low (reference, HR=1) High (HR=11.99 [7.46, 19.25]
Log-rank test: p < 1e-32
Myocardial Infarction
Group Low (reference, HR=1) High (HR=3.73 [2.35, 5.93])
Log-rank test: p = 1.97e-09
Cardiovascular Disease Death
Group Low (reference, HR=1) High (HR=15.69 [7.94, 31.01]
Log-rank test: p < 1e-32
Coronary Heart Disease Death
Group Low (reference, HR=1) High (HR=4.18 [2.51, 6.96])
Log-rank test: p < 1e-32
Stroke
Group Low (reference, HR=1) High (HR=9.05 [4.52, 18.12]
Log-rank test: p = 4.26e-14
Event-free survival probability
Days since baseline
No. at risk
Group High Low


a
β = 0.94, MAE = 0.29 bpm
BCG-derived respiratory rate (bpm)
PSG-derived respiratory rate (bpm)
b
β = 0.91, MAE = 1.84 bpm
BCG-derived heart rate (bpm)
PSG-derived heart rate (bpm)
c
Median: 0.83
Median: 0.70
Pearson correlation coefficient
Respiratory
Heartbeat
d
True Sleep Stage
Predicted Sleep Stage
Awake 92.92% 3.23% 2.89% 0.18% 0.78%
N1 21.78% 34.85% 37.06% 0.83% 5.47%
N2 3.65% 8.94% 75.48% 8.97% 2.95%
N3 2.62% 1.72% 46.48% 48.02% 1.15%
REM 5.60% 5.53% 7.39% 0.15% 81.33%
Sleep staging
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
AUROC=0.952
OSA detection
Actual age (years)
Predicted age (years)
Pearson r=0.78
Age prediction
True Positive Rate (Sensitivity)
False Positive Rate (1 - Specificity)
AUROC=0.802
Sex classification
e
SleepFounder
SleepFM-CR
SleepFounder-Scratch
2cCRNN
Sun et al. (2020)
Choi et al. (2024)
DRIVEN
Sleep staging
Cohen's Kappa ↑
OSA detection
AUROC ↑
Age prediction
MAE ↓
Sex classification
AUROC ↑