A Low-Burden Sleep Foundation Model Built on Respiratory and Heartbeat Signals from 780,000+ Hours of Multi-Ethnic Sleep Recordings

Guangkun Nie[a,l,#], Xuesong Chen[b,#], Yichen Wang[b,#], Jingxu Chen[c], Yunhan Shi[d], Jianwen Zhong[e], Weijun Huang[f], Zengrui Jin[g], Fei Lei[h], Leilei Wang[i], Rui Zhao[j], Chao Zhang[g], Kaibing Chen[k], Dongsheng Lv[j], Wei Chen[i], Hongliang Yi[f], Xiangdong Tang[h,*], Shankai Yin[f,*], Yanru Li[d,*], Shenda Hong[a,*], and Yue Leng[m,*]

[a]National Institute of Health Data Science, Peking University, Beijing, China
[b]Five Seasons Medical, Beijing, China
[c]Sleep Medicine Center, Beijing HuiLongGuan Hospital, Peking University, Beijing, China
[d]Department of Otorhinolaryngology Head and Neck Surgery, Beijing Tongren Hospital, Beijing, China
[e]Department of Pediatric Otorhinolaryngology, Shenzhen Hospital of Southern Medical University, Shenzhen, China
[f]Department of Otorhinolaryngology Head and Neck Surgery, Shanghai Sixth People's Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China
[g]Department of Electronic Engineering, Tsinghua University, Beijing, China
[h]Sleep Medicine Center, West China Hospital, Sichuan University, Chengdu, Sichuan, China
[i]Department of Psychiatry, Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou, Zhejiang, China
[j]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
[k]Sleep Medicine Center, Affiliated Hospital of Gansu University of Chinese Medicine, Lanzhou, Gansu, China
[l]School of Intelligence Science and Technology, Peking University, Beijing, China
[m]Department of Psychiatry and Behavioral Sciences, University of California, San Francisco, CA, USA
[#]These authors contributed equally
[*]Correspondence: 2372564613@qq.com, yinshankai@china.com, liyanruru@aliyun.com, hongshenda@pku.edu.cn, yue.leng@ucsf.edu

ABSTRACT

BACKGROUND. Sleep disorders represent a global health challenge, with obstructive sleep apnea alone affecting nearly 85 million Americans, about 80% of whom remain undiagnosed. Polysomnography (PSG), the gold standard for sleep analysis, is unsuitable for routine or home monitoring. Heartbeat and respiratory signals are core vital signs, broadly obtainable across settings. We investigate whether these signals, when leveraged with foundation model techniques, can enable accurate sleep analysis and provide insights into multi-organ health.

METHODS. We introduce SleepFounder, a foundation model leveraging heartbeat and respiratory signals for low-burden sleep monitoring. SleepFounder was developed from more than 780,000 hours of multi-ethnic sleep recordings across 28 cohorts in the United States and China, enabled by the broad accessibility of these signals. We evaluated SleepFounder on downstream tasks including sleep analysis, demographic profiling, and disease diagnosis across multiple physiological systems, using both curated PSG datasets and real-world, multi-center ballistocardiography (BCG) data collected with a custom-developed sleep mat.

RESULTS. SleepFounder achieved strong performance across all evaluated tasks and typically matched baseline models while using only 20% of the training data. Averaged across external

PSG and BCG datasets, it outperformed baselines with Cohen's κ of 0.65 for five-class sleep staging, intraclass correlation coefficient of 0.85 for apnea-hypopnea index estimation, mean absolute error of 6.90 years for age prediction, and area under the receiver operating characteristic curve (AUROC) of 0.85 for sex classification. In addition, SleepFounder demonstrated broad diagnostic potential: in the held-out test set, it achieved AUROCs of 0.89 for gastroesophageal reflux disease, 0.88 for heart failure, 0.83 for high cholesterol, 0.81 for coronary heart disease, chronic pain, and bipolar disorder, and 0.77 for atrial fibrillation and flutter. These findings were further validated in temporally and geographically external cohorts across five datasets.

CONCLUSIONS. Cardiorespiratory signals encode rich, system-wide physiological information, enabling both conventional sleep analysis and broader health assessment. Building on this insight, SleepFounder demonstrates that low-burden monitoring based solely on respiration and heartbeat can deliver accurate sleep care, highlighting its potential for scalable real-world deployment.

KEYWORDS

Sleep Monitoring, Foundation Model, cardiorespiratory Signals, Polysomnography, Ballistocardiography, Real-World.

INTRODUCTION

Sleep is essential for overall health, with poor sleep quality linked to increased risks of cardiovascular diseases[1–4], 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 diagnosing sleep disorders, 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 routine or home-based monitoring[14,15]. There is an urgent need for cost-effective, low-burden solutions for reliable sleep monitoring.

Recent studies have investigated the use of reduced subsets of PSG channels to enable low-burden sleep monitoring. In particular, deep learning methods applied to cardiorespiratory signals have achieved strong performance in key sleep analysis tasks, including sleep staging[16–19] and OSA severity/apnea–hypopnea index (AHI) estimation[20–23], while minimizing user burden. These signals, which capture the dynamics of the cardiovascular and respiratory systems, are closely associated with sleep patterns and overall physiological states. Moreover, because they are routinely collected in both clinical PSG settings and home-based devices[19,23], they generate large-scale datasets that are well suited for developing advanced deep learning approaches, particularly foundation models. Such models, which leverage large-scale pretraining and generalize across diverse downstream tasks, have shown considerable promise in the medical domain, including computational pathology[24,25], echocardiography[26], and physiological signal analysis[27–31]. Building on these insights, we hypothesize that large-scale cardiorespiratory data, combined with foundation model techniques, can support accurate sleep monitoring and extend to broader multi-organ health analysis.

In this study, we present SleepFounder, a foundation model for accurate, low-burden sleep monitoring that relies exclusively on heartbeat and respiratory signals. SleepFounder was developed using large-scale, multi-ethnic datasets comprising over 780,000 hours of sleep recordings

from 28 cohorts in the United States and China. During pretraining, it demonstrated strong performance on physiology-inspired pretext tasks, including EEG spectrogram reconstruction and oxygen desaturation prediction, thereby validating that respiratory and heartbeat signals capture critical physiological information essential for sleep monitoring. Unlike prior foundation models such as SleepFM[31], which depend on resource-intensive PSG data, SleepFounder's reliance on two widely accessible signals enables flexible, low-burden sleep analysis across diverse devices, supporting applications in both clinical and home settings.

We evaluated SleepFounder across a wide range of downstream tasks, including sleep analysis, demographic profiling, and disease diagnosis spanning multiple physiological systems. Evaluations were performed using both curated PSG datasets and real-world, multi-center BCG data collected with our previously developed sleep mat[32]. Across all tasks, SleepFounder consistently achieved state-of-the-art performance and often matched or exceeded baseline models while using only 20% of the training data. Compared with baseline models, it achieved the best performance in five-class sleep staging (Cohen's κ = 0.65), apnea-hypopnea index estimation (intraclass correlation coefficient [ICC] = 0.85), age prediction (mean absolute error [MAE] = 6.90 years), and sex classification (area under the receiver operating characteristic curve [AUROC] = 0.85). In addition, SleepFounder demonstrated strong diagnostic utility across a wide range of diseases. In the held-out test set, it achieved AUROCs of 0.89 for gastroesophageal reflux disease (GERD), 0.88 for heart failure, 0.83 for high cholesterol, 0.81 for coronary heart disease (CHD), chronic pain, and bipolar disorder, and 0.77 for atrial fibrillation and flutter (AF), with confirmation through further temporally and geographically external cohorts. These findings were further confirmed in temporally external and geographically independent cohorts, underscoring the robustness and generalizability of the model. Collectively, these results demonstrate that easily obtainable cardiorespiratory signals during sleep capture rich, system-wide physiological information and can support accurate assessment of both sleep health and broader systemic health. Leveraging this capability, SleepFounder emerges as a promising foundation model for scalable, low-burden sleep monitoring.

RESULTS

Study Population and Method Overview

The development and evaluation of SleepFounder relied on large-scale, multi-center, and demographically diverse datasets. In total, 28 datasets were included, spanning multiple centers in the United States and China, with baseline demographic characteristics summarized in Table 1. For model pretraining, 23 datasets (n = 50,223) were utilized, incorporating physiologically motivated pretext tasks including EEG spectrogram reconstruction and oxygen desaturation prediction. Model evaluation was conducted on four independent PSG datasets, namely Sleep Heart Health Study (SHHS, n = 5,795) and Wisconsin Sleep Cohort (WSC, n = 1,123) from the United States, and Shanghai Sixth People's Hospital (SSPH, n = 6,609) and Sir Run Run Shaw Hospital (SRRSH, n = 4,014) from China. In addition, real-world validation was performed using multi-center BCG datasets (n = 2,179) collected with a previously developed sleep mat system from three independent institutions in China: The Affiliated Hospital of Gansu University of Chinese Medicine, Shanghai Sixth People's Hospital, and Inner Mongolia Brain Hospital. These BCG datasets enabled assessment of SleepFounder's performance for low-burden, home-based sleep monitoring.

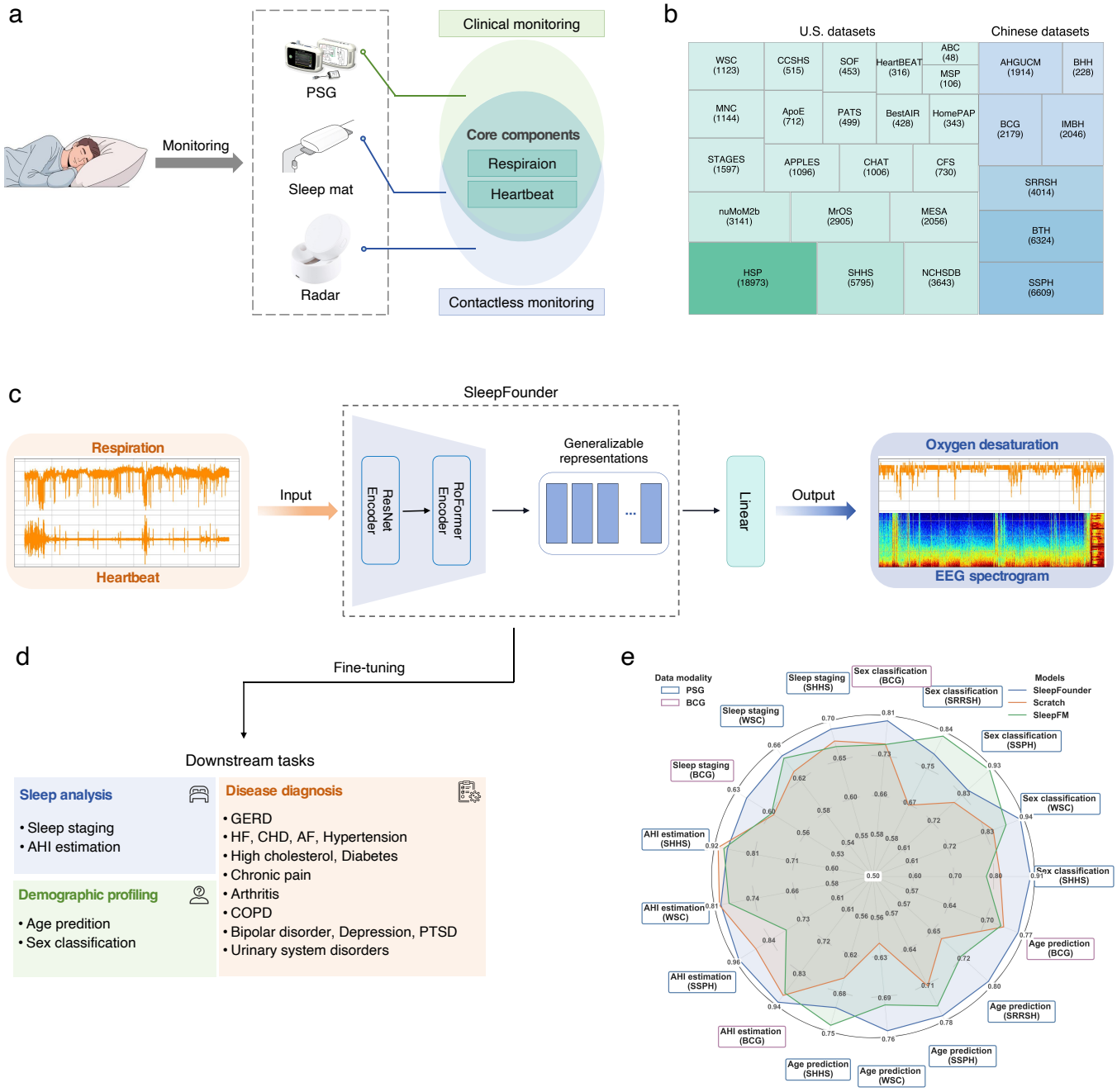


Figure 1: Overview of the proposed SleepFounder model. (a) Clinical and home-based monitoring devices, including PSG, sleep mats, and radar systems, capture core cardiorespiratory signals. (b) Model development incorporated large-scale, multi-ethnic datasets (n = 69,943) from the United States and China. Each rectangle corresponds to one dataset, with its area scaled to the cube root of the sample size to improve visualization. Color hue denotes country of origin (green for the United States, blue for China), and darker shades indicate larger sample sizes. (c) SleepFounder integrates a ResNet-based feature extractor with a RoFormer encoder and is pretrained on respiratory and heartbeat signals using physiologically inspired objectives, including EEG spectrogram reconstruction and oxygen desaturation prediction, to learn robust and generalizable representations. (d) The pretrained model can be fine-tuned for diverse downstream tasks, including sleep analysis, demographic profiling, and disease diagnosis across multiple physiological systems. (e) Radar plots compare SleepFounder with baseline models across different task-dataset pairs.

Table 1: Summary statistics of the datasets used for pretraining and evaluation. For continuous variables, the mean and standard deviation (mean $\pm$ 2 SD) are reported. For categorical variables, the number of subjects and corresponding percentages are provided.

Dataset	No. Subjects	Age	Male	Race				
				White	Asian	Black or African American	Hispanic	Other
Pretraining								
HSP	18973	53.04 ± 33.64	13544 (57.27%)	–	–	–	–	–
NCHSDB	3643	8.78 ± 11.90	–	–	–	–	–	–
nuMoM2b	3141	27.00 ± 10.99	0 (0.00%)	3269 (62.39%)	192 (3.66%)	592 (11.30%)	945 (18.03%)	242 (4.62%)
MrOS	2905	77.57 ± 11.20	3930 (100.00%)	3540 (91.57%)	134 (3.47%)	140 (3.62%)	0 (0.00%)	52 (1.35%)
MESA	2056	69.37 ± 18.23	954 (46.40%)	743 (36.14%)	250 (12.16%)	572 (27.82%)	491 (23.88%)	0 (0.00%)
STAGES	1597	46.00 ± 30.07	768 (48.82%)	1228 (78.27%)	164 (10.45%)	63 (4.02%)	0 (0.00%)	114 (7.27%)
MNC	1144	–	–	–	–	–	–	–
APPLES	1096	51.55 ± 24.45	720 (65.69%)	836 (76.28%)	61 (5.57%)	99 (9.03%)	77 (7.03%)	23 (2.10%)
CHAT	1006	6.95 ± 2.91	181 (48.40%)	253 (37.82%)	0 (0.00%)	331 (49.48%)	0 (0.00%)	85 (12.71%)
CFS	730	41.39 ± 38.78	329 (45.07%)	302 (41.37%)	0 (0.00%)	406 (55.62%)	0 (0.00%)	22 (3.01%)
ApoE	712	45.70 ± 27.24	421 (59.13%)	–	–	–	–	–
CCSHS	515	17.72 ± 0.87	260 (50.49%)	307 (59.61%)	0 (0.00%)	185 (35.92%)	0 (0.00%)	23 (4.47%)
PATS	499	7.04 ± 4.67	378 (51.71%)	488 (67.13%)	13 (1.79%)	184 (25.31%)	0 (0.00%)	42 (5.78%)
SOF	453	82.85 ± 6.28	0 (0.00%)	415 (91.61%)	0 (0.00%)	38 (8.39%)	0 (0.00%)	0 (0.00%)
BestAIR	428	64.79 ± 14.19	170 (67.73%)	229 (91.97%)	6 (2.41%)	14 (5.62%)	0 (0.00%)	0 (0.00%)
HomePAP	343	46.36 ± 23.66	239 (57.87%)	298 (73.76%)	5 (1.24%)	89 (22.03%)	0 (0.00%)	12 (2.97%)
HeartBEAT	316	63.21 ± 14.52	430 (73.38%)	474 (81.16%)	12 (2.05%)	75 (12.84%)	0 (0.00%)	23 (3.94%)
MSP	106	26.76 ± 11.94	0 (0.00%)	8 (7.62%)	2 (1.90%)	95 (90.48%)	0 (0.00%)	0 (0.00%)
ABC	48	–	–	–	–	–	–	–
BTH	6324	–	–	–	–	–	–	–
BHH	228	43.60 ± 35.52	48 (50.53%)	0 (0.00%)	95 (100.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
AHGUCM	1914	46.71 ± 32.05	1250 (65.31%)	0 (0.00%)	1914 (100.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
IMBH	2046	52.02 ± 30.56	704 (34.41%)	0 (0.00%)	2046 (100.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Evaluation								
SHHS	5795	64.54 ± 22.33	3983 (47.19%)	7206 (85.38%)	0 (0.00%)	695 (8.23%)	0 (0.00%)	539 (6.39%)
WSC	1123	59.82 ± 16.98	1385 (53.89%)	2463 (95.84%)	23 (0.89%)	36 (1.40%)	26 (1.01%)	22 (0.86%)
SSPH	6609	41.84 ± 27.80	4358 (79.04%)	0 (0.00%)	5514 (100.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
SRRSH	4014	49.89 ± 31.55	1808 (43.01%)	0 (0.00%)	4205 (100.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
BCG	2179	41.28 ± 27.10	1127 (51.72%)	0 (0.00%)	2179 (100.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)

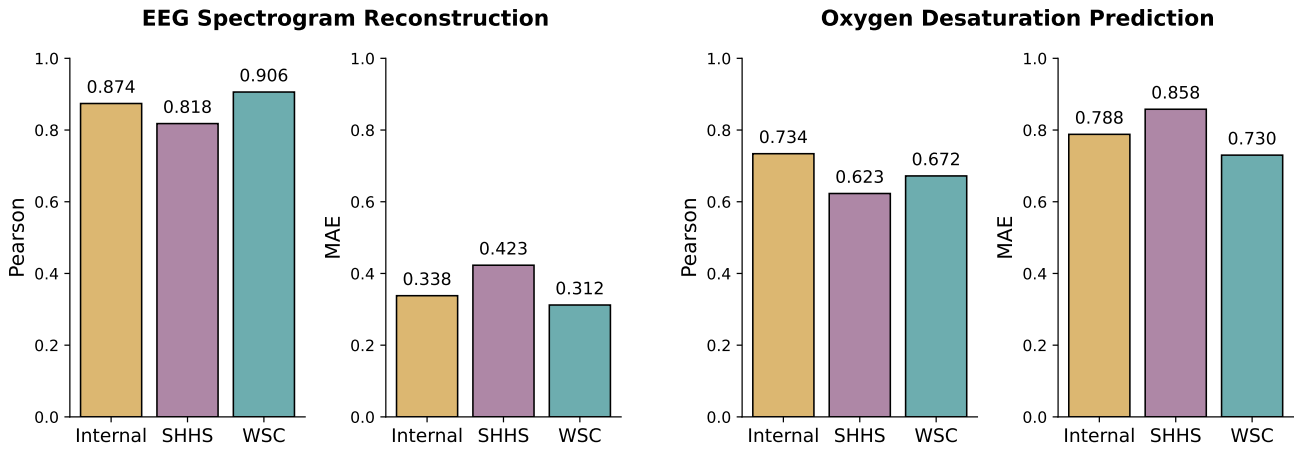


Figure 2: Representative examples of SleepFounder's pretraining performance on two pretext tasks: (a) EEG spectrogram reconstruction and (b) oxygen desaturation prediction, with an additional large figure at the bottom.

SleepFounder Learns Physiologically Relevant Features during Pretraining

SleepFounder was pretrained on physiologically relevant pretext tasks, using heartbeat and respiratory signals to generate EEG spectrograms and oxygen desaturation curves. Quantitative evaluation across our internal dataset and two external cohorts (SHHS and WSC) confirmed strong generalization of the pretrained representations (Figure 2). For EEG spectrogram reconstruction, the model achieved high Pearson correlations of 0.874 (internal), 0.818 (SHHS), and 0.906 (WSC). Performance was similarly consistent for oxygen desaturation prediction, with correlations of 0.734 (internal), 0.623 (SHHS), and 0.672 (WSC), respectively. The consistently low MAE values further validate the physiological relevance and robustness of SleepFounder's learned features.

SleepFounder Demonstrates Superior Performance on Sleep Analysis and Demographic Profiling

We validated SleepFounder on four external datasets, namely SHHS, WSC, SRRSH, and SSPH, from the United States and China that were excluded from pretraining. These datasets included both conventional sleep analysis tasks (five-class sleep staging and AHI estimation) and emerging sleep health profiling tasks (age prediction and sex classification).

Across all four tasks, SleepFounder consistently achieved state-of-the-art performance (Figure 3a). In internal validation, it reached Cohen's κ of 0.693 (95% CI: 0.687-0.698) for sleep staging, ICC of 0.951 (0.943-0.962) for AHI estimation, MAE of 4.898 (4.747-5.045) years for age prediction, and AUROC of 0.943 (0.937-0.950) for sex classification. Each of these values represented the best performance among compared models. Importantly, in external validation across four independent cohorts, SleepFounder outperformed baselines in 9 of 13 task-dataset pairs, with average performance of Cohen's κ = 0.649 for sleep staging, ICC = 0.877 for AHI estimation, MAE = 6.898 years for age prediction, and AUROC = 0.852 for sex classification. These results highlight its strong generalizability across populations, modalities, and recording environments. When compared with the most competitive PSG-based foundation model SleepFM, SleepFounder achieved superior average performance in all four tasks, with particularly notable improvements in AHI estimation (+7.02%) and age prediction (+8.07%). In addition, relative to end-to-end

counterparts, SleepFounder demonstrated superior performance in 11 of 13 external evaluations, highlighting the robustness of its large-scale pretraining strategy. Taken together, these findings establish SleepFounder as a reliable and generalizable foundation model for both traditional and emerging sleep health applications.

We further evaluated the data efficiency of SleepFounder. 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 2%-50% of the labels (Fig. 3b and Fig. 3c). This efficiency was most evident in sleep staging and AHI estimation, where comparable accuracy could be achieved with merely 2%-15% of the data. Strikingly, in extreme low-resource settings, SleepFounder still delivered clinically meaningful results. For example, using only 20 training samples, it reached κ values of 0.663 and 0.595 in SHHS and WSC for sleep staging, and ICC values of 0.886 and 0.939 in SHHS and SSPH for AHI estimation. These findings demonstrate that the representations learned during pretraining are highly transferable and closely aligned with sleep physiology. As a result, SleepFounder can be effectively adapted to new datasets and clinical settings even under severe data constraints, highlighting its potential to enable equitable and scalable sleep health applications.

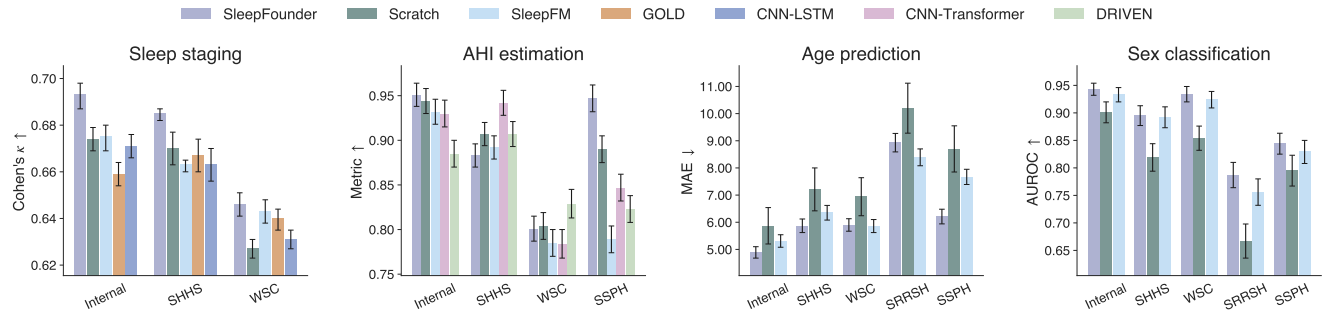
SleepFounder Exhibits Substantial Clinical Value in Multi-system, Multi-disease Diagnosis

Sleep is a fundamental physiological state in which multiple systems are dynamically regulated and interact. Building on this premise, we evaluated the ability of SleepFounder to extend beyond sleep characterization toward comprehensive health assessment using respiratory and heartbeat signals collected during sleep. For this evaluation, SleepFounder was fine-tuned on the pretraining dataset with consolidated diagnostic labels spanning diverse physiological systems and disease categories, and subsequently tested on a held-out internal test set, a temporally external HSP cohort, and four independent external cohorts (SHHS, WSC, SSPH, and SRRSH), none of which were used during pretraining or fine-tuning.

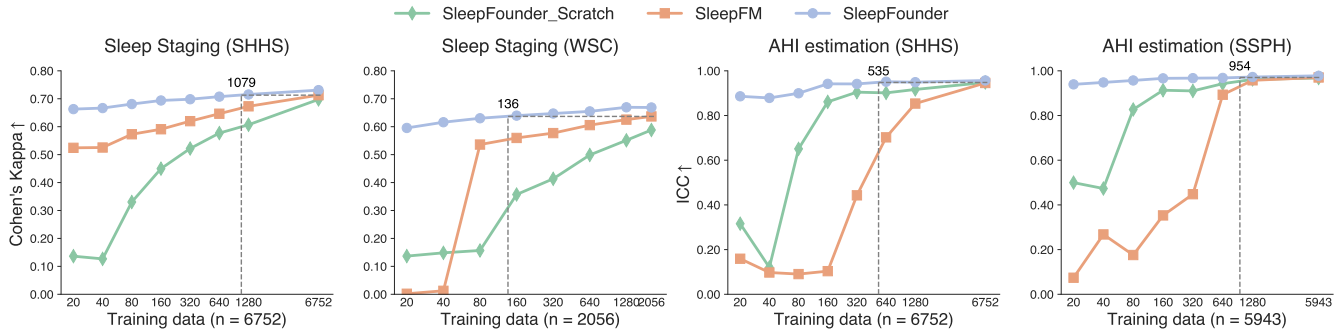
Across seven physiological systems and sixteen disease categories, SleepFounder achieved clinically meaningful diagnostic performance (Figure 4, Table S2). On the held-out internal test set, AUROCs were 0.886 for GERD in the digestive system; 0.881 for heart failure, 0.814 for CHD, 0.776 for cerebrovascular disease, 0.768 for AF, and 0.721 for hypertension in the cardiovascular system; 0.830 for high cholesterol and 0.785 for diabetes in the endocrine and metabolic system; 0.810 for chronic pain and 0.736 for arthritis in the musculoskeletal system; 0.744 for chronic obstructive pulmonary disease (COPD) in the respiratory system; 0.811 for bipolar disorder, 0.727 for depression, 0.724 for post-traumatic stress disorder (PTSD), and 0.717 for anxiety in mental and behavioral disorders; and 0.737 for urologic or kidney problems in the genitourinary system. These results highlight broad and clinically relevant coverage across major physiological systems.

Performance generalized well to external cohorts. In the temporally external HSP cohort, diagnostic performance remained robust across multiple systems, with AUROCs ranging from 0.688 to 0.851 across cardiovascular, metabolic, musculoskeletal, respiratory, mental, and genitourinary conditions. In independent cohorts from both the United States and China, SleepFounder consistently retained discriminative ability: for example, diabetes was predicted with AUROCs of 0.722 in SHHS, 0.765 in WSC, 0.731 in SSPH, and 0.741 in SRRSH; CHD with 0.718 in SHHS, 0.760 in WSC, and 0.749 in SRRSH; and hypertension with 0.701 in SHHS, 0.707 in WSC, 0.672 in SSPH, and 0.733 in SRRSH. Additional robustness was observed in respiratory outcomes, with COPD achieving AUROCs of 0.722 in SHHS and 0.740 in WSC, and in musculoskeletal outcomes, with arthritis reaching 0.680 in WSC.

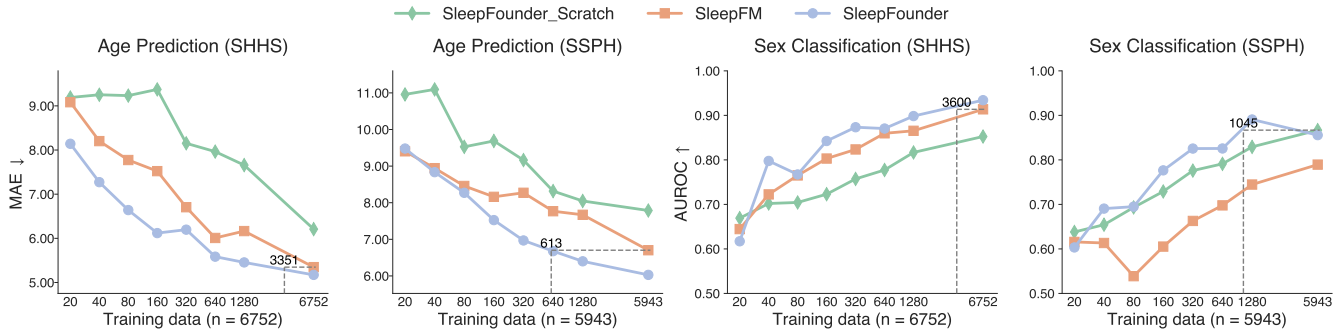
Collectively, these findings demonstrate that SleepFounder delivers broad, system-level



(a) Internal and external validation results, showing that SleepFounder consistently outperformed baseline models.



(b) Under limited data conditions, SleepFounder demonstrated superior data efficiency in sleep staging and AHI estimation.



(c) Under the same constraints, SleepFounder also achieved the best data efficiency in demographic profiling tasks, including age prediction and sex classification.

Figure 3: Comprehensive evaluation of SleepFounder across downstream tasks.

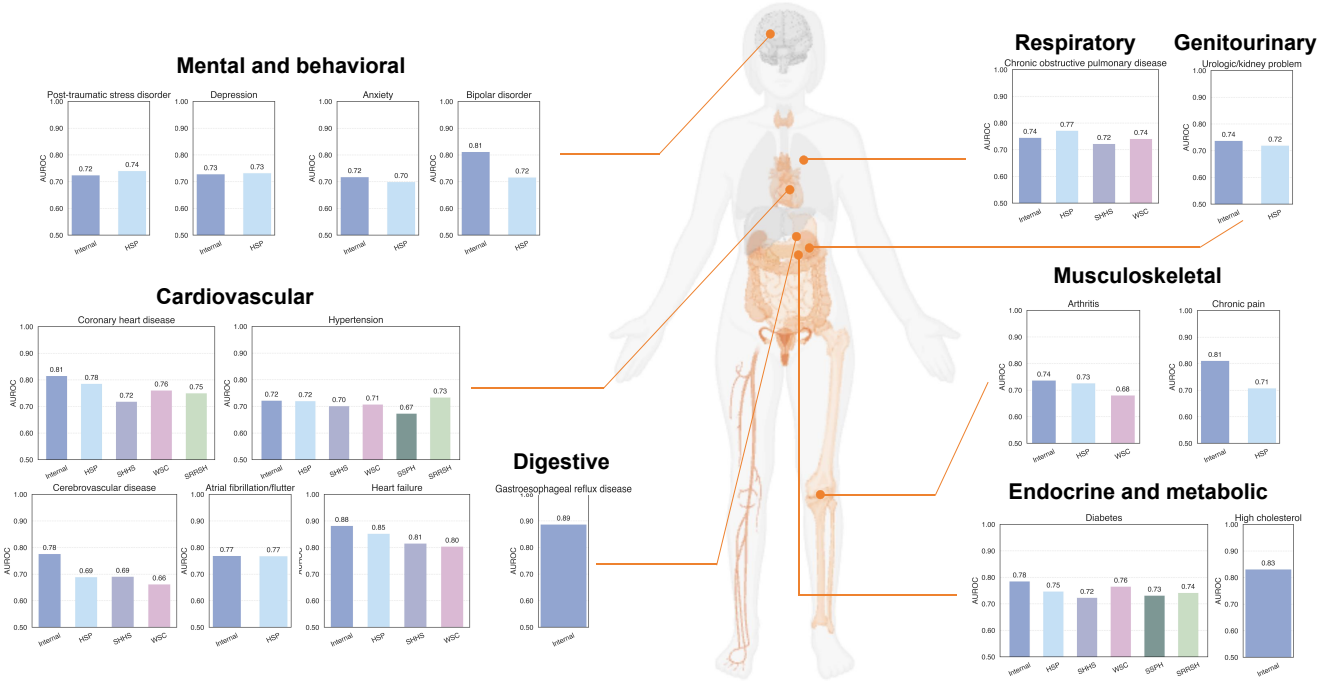


Figure 4: Diagnostic performance (AUROC) of SleepFounder across diseases and cohorts.

diagnostic utility and maintains consistent performance across internal, temporal, and fully external cohorts, underscoring the value of sleep-derived cardiorespiratory signals for scalable and low-burden integrative health assessment.

SleepFounder Achieves Robust Performance on Real-world BCG Device

SleepFounder was pretrained on respiratory and heartbeat 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. As a representative example, we analyzed BCG signals collected with our previously developed sleep mat system. For the purpose of assessing cross-modal agreement, we used recordings in which PSG and BCG were acquired simultaneously.As shown in Figure 5a and 5b, quantitative analyses confirmed strong consistency between BCG- and PSG-derived signals, with respiratory rate showing excellent agreement (Pearson's r = 0.94, MAE = 0.29 breaths/min; and heart rate showing good concordance (Pearson's r = 0.91, MAE = 1.84 beats/min. Distribution analyses further indicated robust cross-modal correlations (Figure 5c). These findings demonstrate that the cardiorespiratory signals used to train SleepFounder are reliably captured by BCG devices, supporting direct application of the model in home-based monitoring.

We then performed external validation to test real-world transfer. Models fine-tuned on PSG were directly evaluated on BCG datasets collected from the three independent centers in China, which were excluded from both pretraining and fine-tuning. Across four downstream tasks, SleepFounder achieved strong accuracy, with Cohen's κ of 0.617 (0.610-0.623) for sleep staging, ICC of 0.926 (0.911-0.938) for AHI estimation, MAE of 7.580 years (7.440-7.730) for age prediction, and AUROC of 0.802 (0.781-0.820) for sex classification (Figure 5d). Compared with multiple baselines, SleepFounder achieved the best performance across all tasks, confirming that SleepFounder can be deployed without additional adaptation on real-world home-based sleep monitoring devices.

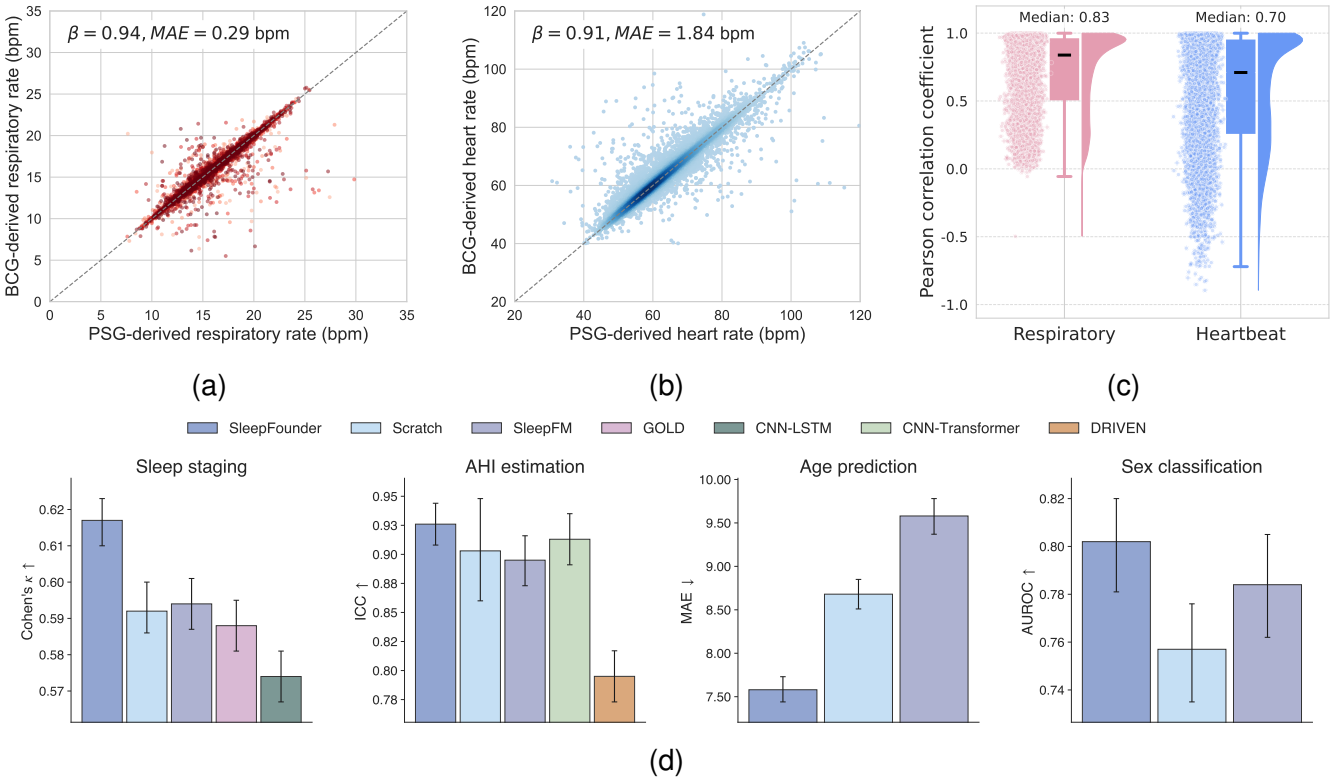


Figure 5: External validation of SleepFounder on real-world BCG data. (a) Density scatter plot of respiratory rates derived from PSG and BCG respiratory signals. (b) Density scatter plot of heart rates derived from PSG and BCG heartbeat signals. (c) Raincloud plots showing the distribution of Pearson correlation coefficients for respiratory and heartbeat signals across PSG–BCG pairs. bpm: breaths/beats per minute; r: Pearson correlation coefficient; MAE: mean absolute error. (d) Performance of SleepFounder compared with baseline models across downstream tasks in the external, multi-center BCG dataset.

DISCUSSION

Sleep disorders affect over one billion people globally[11,33], with the majority remaining undiagnosed and untreated[12,13]. Traditional PSG-based sleep monitoring is resource-intensive and disrupts natural sleep, limiting its scalability. In this study, we introduce SleepFounder, a foundation model that leverages widely accessible heartbeat and respiratory signals, and examine its potential for sleep-centric health monitoring. Compared to prior PSG-based model SleepFM[31], SleepFounder demonstrates superior potential for real-world deployment by relying on practical and widely accessible input modalities.

We evaluated SleepFounder across a range of downstream tasks, including demographic profiling (age prediction and sex classification) and key sleep-related analyses (sleep staging and AHI estimation). To comprehensively validate its real-world applicability, we assessed performance not only on publicly available PSG datasets (e.g., SHHS and WSC) but also on multi-center BCG datasets collected using the sleep mat developed in our prior work. Across all tasks, SleepFounder consistently achieved state-of-the-art results. Remarkably, it often matched or surpassed baseline models trained on the full dataset while requiring only 20% of the data, highlighting its exceptional data efficiency. This property is particularly critical for real-world deployment, where high-quality annotated data are limited and conventional deep learning models struggle to maintain robust accuracy and generalizability. In contrast, SleepFounder's efficient use of scarce data enables strong and reliable performance under practical constraints, underscoring its potential for real-world applications.

Given the close relationship between sleep and overall health, we evaluated the diagnostic potential of SleepFounder across seven physiological systems and sixteen disease categories. Using respiratory and heartbeat signals collected during sleep, the model achieved clinically meaningful performance on the held-out internal test set and maintained robustness in both temporally external and geographically independent cohorts. This consistency underscores the potential of sleep-derived cardiorespiratory signals as a scalable substrate for multi-system disease detection. For example, within the cardiovascular domain, the model reliably discriminated heart failure, CHD, cerebrovascular disease, AF, and hypertension, underscoring its clinical utility in one of the most critical disease areas. A particularly illustrative example was observed in the digestive system, where SleepFounder achieved an AUROC of 0.886 for GERD. This result is clinically interpretable: GERD symptoms often worsen in the supine position during sleep, leading to nocturnal heartburn and regurgitation, which in turn cause frequent arousals, fragmented sleep, and a cascade of cardiorespiratory alterations such as abnormal breathing patterns, reduced heart rate variability, and heightened autonomic responses. By capturing these characteristic cardiorespiratory signatures, SleepFounder successfully identified GERD, further underscoring the link between sleep and systemic health. Taken together, these findings indicate that SleepFounder extends beyond traditional sleep characterization to enable clinically relevant, cross-system disease diagnosis. By maintaining stable performance across internal, temporal, and external cohorts from both the United States and China, the model demonstrates strong generalizability and real-world applicability, supporting the feasibility of a sleep-centered paradigm for integrated health monitoring and disease prediction.

A central finding of this study is that cardiorespiratory signals, particularly respiration and heartbeat, provide substantial value for both sleep analysis and disease diagnosis. Although PSG remains the clinical gold standard for comprehensive sleep assessment, its complexity and intrusiveness constrain scalability. To address this limitation, recent studies have investigated the use of reduced subsets of PSG channels, enabling more accessible sleep monitoring with minimal disruption to natural sleep while maintaining acceptable performance. Many of these efforts emphasize cardiorespiratory signals owing to their accessibility and non-intrusiveness. For instance,

Sridhar et al.[34] employed instantaneous heart rate for automated sleep staging, while Retamales and Hu et al.[22,23] utilized abdominal movement, thoracic movement, and pulse oximetry for sleep apnea assessment. These studies highlight the potential of easily accessible cardiorespiratory signals for accurate sleep monitoring. In this study, we further investigate the feasibility and performance of this approach using foundation models. Through downstream task performance and comparisons with existing works on task-specific PSG channels for sleep staging[31,35,36], AHI estimation[22,23], and disease diagnosis, our findings demonstrate that respiration and heartbeat signals alone are sufficient for reliable sleep analysis. This underscores the potential to shift from complex, intrusive PSG to more accessible, low-burden cardiorespiratory-based approaches, enabling scalable sleep monitoring and expanding sleep assessment from diagnostics to broader health management.

Despite these promising results, several limitations warrant consideration. First, SleepFounder was pretrained exclusively on PSG data. Although we confirmed the high similarity between respiration and heartbeat signals from BCG and PSG and observed robust performance on real-world BCG datasets, the model has not been validated on other sensing modalities, such as radar-based systems. Cross-device differences may thus impact performance, although SleepFounder is still expected to outperform conventional task-specific models. Second, while the model shows potential for disease diagnosis across multiple physiological systems, the underlying datasets have limitations, including reliance on self-reported rather than ICD-confirmed diagnoses. Dataset variability and incomplete disease coverage may introduce label bias, potentially under- or overestimating diagnostic value. Some disease categories could not be fully validated in independent cohorts due to missing labels, and the lack of longitudinal follow-up precludes assessment of the model's utility for disease risk stratification or future outcome prediction. Finally, SleepFounder relies on full-night respiration and heartbeat recordings, which capture comprehensive nocturnal physiological dynamics but limit applicability for real-time or short-duration analyses, such as critical event alerting. Processing full-night data also imposes higher computational and storage demands, adding practical considerations for deployment.

In conclusion, SleepFounder demonstrates the potential of cardiorespiratory-based foundation models for low-burden sleep monitoring. By pretraining on large-scale, multi-ethnic PSG datasets and validating across both clinical PSG and real-world multi-center BCG data, the model achieved robust performance in sleep-related tasks. Our development approach integrates readily accessible yet informative input modalities, physiology-inspired generative self-supervised learning, and rigorous validation across diverse cohorts, providing a unified framework for building accurate and generalizable AI systems for sleep monitoring. These findings highlight the promise of extending foundation models beyond conventional PSG-based analyses toward widely accessible sleep and whole-body health monitoring.

METHODS

Pretraining datasets for SleepFounder

We curated a large-scale, multi-ethnic dataset for pretraining SleepFounder, comprising PSG recordings from 50,223 individuals across 23 datasets in the United States and China. The U.S. component included 19 datasets with recordings from 39,711 participants, covering the HSP dataset and 18 datasets from the NSRR repository. The Chinese component included four datasets with recordings from 10,512 participants. Detailed descriptions of these datasets are provided in the Supplementary Materials.

Human Sleep Project (HSP) The HSP dataset[1] comprises 24,859 PSG recordings from 18,701 patients evaluated at the Massachusetts General Hospital Sleep Division. All recordings were acquired in accordance with American Academy of Sleep Medicine standards and include EEG, electrooculography (EOG), chin electromyography (EMG), respiratory effort, airflow, oxygen saturation, and ECG channels.

National Sleep Research Resource (NSRR) The NSRR[2] provides a comprehensive repository of PSG data, demographic information, and clinical annotations from multiple cohorts. For pretraining, we utilized 18 NSRR datasets, including Apnea, Bariatric Surgery, and CPAP Study (ABC), Sleep Disordered Breathing: ApoE and Lipid Metabolism (ApoE), Apnea Positive Pressure Long-term Efficacy Study (APPLES), Best Apnea Interventions in Research (BestAIR), Cleveland Children's Sleep and Health Study (CCSHS), Cleveland Family Study (CFS), Childhood Adenotonsillectomy Trial (CHAT), Heart Biomarker Evaluation in Apnea Treatment (HeartBEAT), Home Positive Airway Pressure (HomePAP), Multi-Ethnic Study of Atherosclerosis (MESA), Mignot Nature Communications (MNC), MrOS Sleep Study (MrOS), Maternal Sleep in Pregnancy and the Fetus (MSP), NCH Sleep DataBank (NCHSDB), Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-Be (nuMoM2b), Pediatric Adenotonsillectomy Trial of Snoring (PATS), and the Study of Osteoporotic Fractures (SOF).

Beijing Tongren Hospital (BTH) The BTH dataset includes overnight PSG recordings from 6,324 participants collected at the Department of Otorhinolaryngology, Beijing Tongren Hospital. Most participants were patients referred for evaluation of suspected OSA.

Inner Mongolia Brain Hospital (IMBH) The IMBH dataset comprises overnight PSG recordings from 2,046 participants collected at the hospital's sleep center. Participants were patients with psychiatric disorders and comorbid OSA.

The Affiliated Hospital of Gansu University of Chinese Medicine (AHGUCM) The AHGUCM dataset was collected at the Sleep Disorders Department of the Affiliated Hospital of Gansu University of Chinese Medicine between 2017 and 2025, comprising overnight PSG recordings from 1,914 patients. Most patients in this cohort were diagnosed with OSA and insomnia.

[1] https://bdsp.io/content/hsp/2.0/
[2] https://sleepdata.org/

Beijing Huilongguan Hospital (BHH) The BHH dataset was acquired at the Sleep Disorders Center of Beijing Huilongguan Hospital between 2022 and 2025. It consists of overnight PSG recordings from 228 patients with psychiatric disorders, including insomnia, depression, and anxiety.

Data preprocessing for pretraining

Modality-specific techniques were employed to extract heartbeat and respiratory components from PSG and BCG signals. Overnight EEG and SpO_2 recordings were further transformed into spectrograms and desaturation curves, respectively, to serve as reconstruction targets during pretraining. Detailed descriptions of the preprocessing procedures are provided in the Supplementary Materials.

Detail of model architecture and pretraining

SleepFounder consists of a ResNet-based[37] feature extractor followed by an RoFormer encoder[38]. 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 generative 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 $\mathcal{L}_{\text{SpO2}} = \lambda_{\text{Pearson}}\mathcal{L}_{\text{Pearson}} + \lambda_{\text{Wasserstein}}\mathcal{L}_{\text{Wasserstein}}$, where $\lambda_{\text{Pearson}} + \lambda_{\text{Wasserstein}} = 1$. The overall pretraining objective integrates both tasks as a weighted sum, $\mathcal{L}_{\text{total}} = \lambda_{\text{EEG}}\mathcal{L}_{\text{EEG}} + \lambda_{\text{SpO2}}\mathcal{L}_{\text{SpO2}}$, with weights λ_{EEG} and λ_{SpO2} dynamically adjusted according to the relative magnitudes of task-specific losses. This strategy ensures that the learned features preserve both spectral and temporal information critical for downstream sleep monitoring applications.

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.

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 sleep mat[32]. The PSG evaluation datasets included the temporally split HSP dataset and two publicly available NSRR cohorts excluded from pretraining, namely WSC and SHHS, as well as two clinical datasets collected from Zhejiang Sir Run Run Shaw Hospital and Shanghai Sixth People's Hospital in China. The multi-center BCG datasets were collected from Shanghai Sixth People's Hospital, Huaxi Hospital, and a hospital in Inner Mongolia, in parallel with clinical PSG recordings. Detailed information on these datasets is provided in the Supplementary Material.

Sleep Heart Health Study (SHHS) The SHHS[3] dataset is a large-scale, multi-center cohort established to examine the cardiovascular outcomes associated with sleep-disordered breathing. Between 1995 and 1998, over 6,000 middle-aged and older adults underwent baseline overnight PSG assessments, with a follow-up study conducted in the early 2000s for a subset of participants. The dataset further includes extensive longitudinal information on cardiovascular health outcomes, alongside the PSG recordings.

Wisconsin Sleep Cohort (WSC) The WSC[4] dataset originates from a longitudinal study in Wisconsin that began in the late 1980s, designed to investigate sleep-disordered breathing and its health consequences. It contains repeated overnight PSG assessments performed at regular intervals, together with clinical and demographic data.

Sir Run Run Shaw Hospital (SRRSH) The SRRSH dataset refers to data sourced from the Department of Psychiatry at Zhejiang University School of Medicine's Sir Run Run Shaw Hospital, collected between 2018 and 2023. It includes 4,014 usable recordings, primarily focusing on psychiatric disorders such as depression and anxiety. The data from 2021 to 2023 are annotated with corresponding diagnostic results. There is no annotations for sleep stages and AHI events for this dataset.

Shanghai Sixth People's Hospital (SSPH) The SSPH dataset represents data collected from the Department of Otorhinolaryngology at Shanghai Sixth People's Hospital, consisting of 6,609 PSG recordings. It focuses on patients with OSA and includes annotations for sleep stages and respiratory events. It also provides labels for comorbidities including hypertension, diabetes, and dyslipidemia.

BCG The BCG dataset comprises a total of 2,179 PSG recordings, including 945 from the Sleep Medicine Center of West China Hospital, 794 from Shanghai Sixth People's Hospital, and 440 from Inner Mongolia Brain Hospital. For all recordings, PSG and BCG signals were acquired simultaneously, with respiratory events and sleep stages annotated based on the PSG data.

[3] https://sleepdata.org/datasets/shhs
[4] https://sleepdata.org/datasets/wsc

Importantly, this dataset does not overlap in either collection period or patient population with the previously described IMBH and SSPH datasets.

Baseline Methods

To assess the performance of SleepFounder on downstream tasks, we established a set of baseline models for comparison, including self-supervised pretrained model SleepFM and conventional task-specific models for sleeg staging and AHI estimation. For fair comparison, every baseline reproduced in this work was adapted to use the unified 4 Hz breath and heartbeat signal introduced above as inputs. These baselines provide reference points to evaluate the value of our foundation model approach.

Self-supervised pretrained models We included SleepFM[31] as a representative multimodal, self-supervised baseline. SleepFM employs deep encoders with a contrastive learning objective across multiple physiological modalities. For a fair comparison with SleepFounder, we adapted SleepFM by adding a respiration–heartbeat branch consistent with our input setting and architecture. During fine-tuning, only this branch was used to align with SleepFounder's training protocol. Unlike the original SleepFM, which was pretrained primarily on one large-scale PSG dataset, our adaptation relied on a more heterogeneous collection of smaller datasets. While the overall scale was comparable, the greater heterogeneity and stricter modality requirements of SleepFM limited the inclusion of several portable-device PSG datasets, potentially introducing noise and reducing usable data during pretraining.

Baseline models for sleep staging We reproduced two representative convolutional–recurrent architectures. The first was the two-channel network of Goldammer et al.[39], which processes 4-Hz interval-derived streams with stacked 1D convolutions followed by a bidirectional LSTM for epoch-wise staging. Since the original inputs were also sampled at 4 Hz, they were directly replaced by the respiration and heartbeat signals used in this work; the only modification was extending the LSTM's temporal context to full-night sequences (up to 1,535 × 30-s epochs), which improved performance relative to short-window settings. The second baseline was the CNN-LSTM model of Sun et al.[40], originally designed for high-rate ECG (200 Hz) and abdominal/thoracic respiratory effort (10 Hz). In our adaptation, the abdominal CNN encoder was duplicated to process both the preprocessed respiratory signal and the JJ intervals, with signals upsampled to 10 Hz to match the expected input. The original three-stage training procedure was retained, with the only architectural change being an extension of the LSTM temporal window from 28 × 30-s epochs (14 min) to full-night sequences, consistent with SleepFounder.

Baseline models for AHI estimation We reproduced two representative architectures for AHI estimation. The first was DRIVEN[22], a hybrid deep-learning model combining 2D-CNNs with LightGBM. In the original implementation, DRIVEN jointly performed respiratory-event detection and sleep/wake classification to derive total sleep time, but for a fair comparison with SleepFounder, ground-truth sleep labels were used instead. We note that the omission of oximetry, which is included in the original model, may modestly reduce performance relative to the reported results. The second baseline was the CNN-Transformer pipeline of Choi et al.[41], which expects short fixed-length inputs (10-min segments) and produces token-level predictions using a CNN encoder followed by a Transformer. To adapt this model, our 4Hz signals were upsampled to 8 Hz, and the original three-class respiratory-event target (apnea/hypopnea/normal) was simplified to binary event labels. At inference, overlapping 10-min windows with a 30-s hop were applied

across the night, and final per-second probabilities were obtained by averaging predictions from overlapping windows.

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 $\mathbf{h}^{\text{agg}} = \sum_{l=1}^{L} \alpha_l \mathbf{h}^{(l)}$, 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, except for the HSP dataset, which was split temporally. External test datasets were evaluated in their entirety without further partitioning. All experiments were performed on eight NVIDIA H20 GPUs.

Multi-system, multi-disease diagnosis

For disease diagnosis tasks, diagnostic labels were derived from both clinical records and participant self-reports. Whenever available, International Classification of Diseases (ICD)-based clinical diagnoses were prioritized as the primary reference standard. Self-reported diagnoses were used as supplementary labels when ICD-based information was not available. This hierarchical approach ensured that clinical diagnoses were preferentially adopted while maintaining broader coverage of disease outcomes across cohorts, thereby enhancing the reliability and inclusiveness of the diagnostic evaluation.

Evaluation metrics

Task-specific evaluation metrics were used according to the prediction target. For age prediction, performance was assessed using MAE, which quantifies the average absolute difference between predicted and true values. For sex classification and disease diagnosis, AUROC was reported, summarizing model discrimination across all classification thresholds (1 indicates perfect discrimination, 0.5 represents random guessing). For sleep staging, we used Cohen's κ coefficient, which measures agreement between predicted and reference labels while accounting for chance. For AHI estimation, agreement between predicted and true values was quantified using the ICC, with values closer to 1 indicating stronger concordance, and pearson correlation coeeficient was used for a aulixry metric.

ACKNOWLEDGMENTS

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. 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).

DECLARATION OF INTERESTS

The authors declare no competing interests.

References

1. Ge, L., Guyatt, G., Tian, J., Pan, B., Chang, Y., Chen, Y., Li, H., Zhang, J., Li, Y., Ling, J. et al. (2019). Insomnia and risk of mortality from all-cause, cardiovascular disease, and cancer: Systematic review and meta-analysis of prospective cohort studies. Sleep medicine reviews *48*, 101215.

2. Zheng, B., Yu, C., Lv, J., Guo, Y., Bian, Z., Zhou, M., Yang, L., Chen, Y., Li, X., Zou, J. et al. (2019). Insomnia symptoms and risk of cardiovascular diseases among 0.5 million adults: a 10-year cohort. Neurology *93*, e2110–e2120.

3. Floras, J.S. (2018). Sleep apnea and cardiovascular disease: an enigmatic risk factor. Circulation research *122*, 1741–1764.

4. Yeghiazarians, Y., Jneid, H., Tietjens, J.R., Redline, S., Brown, D.L., El-Sherif, N., Mehra, R., Bozkurt, B., Ndumele, C.E., Somers, V.K. et al. (2021). Obstructive sleep apnea and cardiovascular disease: a scientific statement from the american heart association. Circulation *144*, e56–e67.

5. Chasens, E.R., Imes, C.C., Kariuki, J.K., Luyster, F.S., Morris, J.L., DiNardo, M.M., Godzik, C., Jeon, B., and Yang, K. (2021). Sleep and metabolic syndrome. The Nursing Clinics of North America *56*, 203.

6. Che, T., Yan, C., Tian, D., Zhang, X., Liu, X., and Wu, Z. (2021). The association between sleep and metabolic syndrome: a systematic review and meta-analysis. Frontiers in Endocrinology *12*, 773646.

7. Pearson, O., Uglik-Marucha, N., Miskowiak, K.W., Cairney, S.A., Rosenzweig, I., Young, A.H., and Stokes, P.R. (2023). The relationship between sleep disturbance and cognitive impairment in mood disorders: a systematic review. Journal of affective disorders *327*, 207–216.

8. Maggi, G., Trojano, L., Barone, P., and Santangelo, G. (2021). Sleep disorders and cognitive dysfunctions in parkinson's disease: a meta-analytic study. Neuropsychology review pp. 1–40.

9. Palagini, L., Hertenstein, E., Riemann, D., and Nissen, C. (2022). Sleep, insomnia and mental health. Journal of sleep research *31*, e13628.

10. Kim, J.Y., Ko, I., and Kim, D.K. (2019). Association of obstructive sleep apnea with the risk of affective disorders. JAMA Otolaryngology–Head & Neck Surgery *145*, 1020–1026.

11. Benjafield, A.V., Ayas, N.T., Eastwood, P.R., Heinzer, R., Ip, M.S., Morrell, M.J., Nunez, C.M., Patel, S.R., Penzel, T., Pépin, J.L. et al. (2019). Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. The Lancet respiratory medicine *7*, 687–698.

12. Watson, N., Yu, K., Campbell, D., Bloudek, L., Yee, J., Cronin, J., Azofra, A.S., and Boskovic, N. (2025). 0637 prevalence and unmet need of obstructive sleep apnea in the united states. Sleep *48*, A278–A278.

13. Faria, A., Allen, A.H., Fox, N., Ayas, N., and Laher, I. (2021). The public health burden of obstructive sleep apnea. Sleep Science *14*, 257–265.

14. Li, X., Li, M., Tian, M., Liu, Q., Zhou, X., Liu, H., Li, R., Li, Z., Dong, H., Jia, L. et al. (2025). Exploring the potential of a new wearable sleep monitoring device for clinical application. Biomedical Signal Processing and Control *99*, 106856.

15. Kwon, S., Kim, H., and Yeo, W.H. (2021). Recent advances in wearable sensors and portable electronics for sleep monitoring. Iscience *24*.

16. Du, Y., Gu, J., Duan, S., Trueb, J., Tzavelis, A., Shin, H.S., Arafa, H., Li, X., Huang, Y., Carr, A.N. et al. (2025). A skin-interfaced wireless wearable device and data analytics approach for sleep-stage and disorder detection. Proceedings of the National Academy of Sciences *122*, e2501220122.

17. Bakker, J.P., Ross, M., Vasko, R., Cerny, A., Fonseca, P., Jasko, J., Shaw, E., White, D.P., and Anderer, P. (2021). 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 *17*, 1343–1354.

18. Anderer, P., Ross, M., Cerny, A., Radha, M., and Fonseca, P. (2020). 0436 deep learning for scoring sleep based on signals available in home sleep apnea test studies: Cardiorespiratory sleep staging. Sleep *43*, A167.

19. Cerina, L., Overeem, S., Papini, G.B., van Dijk, J.P., Vullings, R., van Meulen, F., Ross, M., Cerny, A., Anderer, P., and Fonseca, P. (2024). A sleep stage estimation algorithm based on cardiorespiratory signals derived from a suprasternal pressure sensor. Journal of Sleep Research *33*, e14015.

20. Ross, M., Fonseca, P., Overeem, S., Vasko, R., Cerny, A., Shaw, E., and Anderer, P. (2023). Autonomic arousal detection and cardio-respiratory sleep staging improve the accuracy of home sleep apnea tests. Frontiers in Physiology *14*, 1254679.

21. Fonseca, P., Ross, M., Cerny, A., Anderer, P., Schipper, F., Grassi, A., van Gilst, M., and Overeem, S. (2024). Estimating the severity of obstructive sleep apnea using ecg, respiratory effort and neural networks. IEEE Journal of Biomedical and Health Informatics *28*, 3895–3906.

22. Retamales, G., Gavidia, M.E., Bausch, B., Montanari, A.N., Husch, A., and Goncalves, J. (2024). Towards automatic home-based sleep apnea estimation using deep learning. npj Digital Medicine *7*, 144.

23. Hu, S., Liu, J., Wang, Y., Fu, C., Zhu, J., Yu, H., and Yang, C. (2025). Transparent artificial intelligence-enabled interpretable and interactive sleep apnea assessment across flexible monitoring scenarios. Nature Communications *16*, 7548.

24. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M. et al. (2024). Towards a general-purpose foundation model for computational pathology. Nature Medicine *30*, 850–862.

25. Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., Zimmermann, E., Hall, J., Tenenholtz, N., Fusi, N. et al. (2024). A foundation model for clinical-grade computational pathology and rare cancers detection. Nature medicine *30*, 2924–2935.

26. Christensen, M., Vukadinovic, M., Yuan, N., and Ouyang, D. (2024). Vision–language foundation model for echocardiogram interpretation. Nature Medicine *30*, 1481–1488.

27. Abbaspourazad, S., Elachqar, O., Miller, A., Emrani, S., Nallasamy, U., and Shapiro, I. (). Large-scale training of foundation models for wearable biosignals. In The Twelfth International Conference on Learning Representations.

28. Tian, Y., Li, Z., Jin, Y., Wang, M., Wei, X., Zhao, L., Liu, Y., Liu, J., and Liu, C. (2024). Foundation model of ecg diagnosis: Diagnostics and explanations of any form and rhythm on ecg. Cell Reports Medicine *5*.

29. Li, J., Aguirre, A.D., Junior, V.M., Jin, J., Liu, C., Zhong, L., Sun, C., Clifford, G., Brandon Westover, M., and Hong, S. (2025). An electrocardiogram foundation model built on over 10 million recordings. NEJM AI *2*, AIoa2401033.

30. Li, J., Goldenholz, D.M., Alkofer, M., Sun, C., Nascimento, F.A., Halford, J.J., Dean, B.C., Galanti, M., Struck, A.F., Greenblatt, A.S. et al. (2025). Expert-level detection of epilepsy markers in eeg on short and long timescales. NEJM AI *2*, AIoa2401221.

31. Thapa, R., Kjær, M.R., He, B., Covert, I., Moore, H., Hanif, U., Ganjoo, G., Westover, M.B., Jennum, P., Brink-Kjær, A. et al. (2025). A multimodal sleep foundation model developed with 500k hours of sleep recordings for disease predictions. medRxiv.

32. Li, S., Chen, Y., Chen, X., Gao, R., Zhang, Y., Yu, C., Li, Y., Ye, Z., Huang, W., Yi, H. et al. (2024). Sleepnetzero: Zero-burden zero-shot reliable sleep staging with neural networks based on ballistocardiograms. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies *8*, 1–25.

33. Benjafield, A.V., Kuniyoshi, F.H.S., Malhotra, A., Martin, J.L., Morin, C.M., Maurer, L.F., Cistulli, P.A., Pépin, J.L., Wickwire, E.M. et al. (2025). Estimation of the global prevalence and burden of insomnia: a systematic literature review-based analysis. Sleep Medicine Reviews *82*, 102121.

34. Sridhar, N., Shoeb, A., Stephens, P., Kharbouch, A., Shimol, D.B., Burkart, J., Ghoreyshi, A., and Myers, L. (2020). Deep learning for automated sleep staging using instantaneous heart rate. NPJ digital medicine *3*, 106.

35. Fox, B., Jiang, J., Wickramaratne, S., Kovatch, P., Suarez-Farinas, M., Shah, N.A., Parekh, A., and Nadkarni, G.N. (2025). A foundational transformer leveraging full night, multichannel sleep study data accurately classifies sleep stages. Sleep pp. zsaf061.

36. Perslev, M., Darkner, S., Kempfner, L., Nikolic, M., Jennum, P.J., and Igel, C. (2021). U-sleep: resilient high-frequency sleep staging. NPJ digital medicine *4*, 72.

37. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778.

38. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. (2024). Roformer: Enhanced transformer with rotary position embedding. Neurocomputing *568*, 127063.

39. Goldammer, M., Zaunseder, S., Brandt, M.D., Malberg, H., and Gräßer, F. (2022). Investigation of automated sleep staging from cardiorespiratory signals regarding clinical applicability and robustness. Biomedical Signal Processing and Control *71*, 103047. URL: https://www.sciencedirect.com/science/article/pii/S1746809421006443. doi: https://doi.org/10.1016/j.bspc.2021.103047.

40. Sun, H., Ganglberger, W., Panneerselvam, E., Leone, M.J., Quadri, S.A., Goparaju, B., Tesh, R.A., Akeju, O., Thomas, R.J., and Westover, M.B. (2019). Sleep staging from electrocardiography and respiration with deep learning. Sleep *43*, zsz306. URL: https://doi.org/10.1093/sleep/zsz306. doi: 10.1093/sleep/zsz306. arXiv:https://academic.oup.com/sleep/article-pdf/43/7/zsz306/33460809/zsz306.pdf.

41. Choi, J.W., Koo, D.L., Kim, D.H., Nam, H., Lee, J.H., Hong, S.N., and Kim, B. (2024). A novel deep learning model for obstructive sleep apnea diagnosis: hybrid cnn-transformer approach for radar-based detection of apnea-hypopnea events. Sleep *47*, zsae184. URL: https://doi.org/10.1093/sleep/zsae184. doi: 10.1093/sleep/zsae184. arXiv:https://academic.oup.com/sleep/article-pdf/47/12/zsae184/60774106/zsae184.pdf.

42. Makowski, D., Pham, T., Lau, Z.J., Brammer, J.C., Lespinasse, F., Pham, H., Schölzel, C., and Chen, S.A. (2021). Neurokit2: A python toolbox for neurophysiological signal processing. Behavior research methods pp. 1–8.

43. Brüser, C., Winter, S., and Leonhardt, S. (2013). Robust inter-beat interval estimation in cardiac vibration signals. Physiological measurement *34*, 123.

44. Massaro, S., and Pecchia, L. (2019). Heart rate variability (hrv) analysis: A methodology for organizational neuroscience. Organizational research methods *22*, 354–393.

45. Thomson, W. (1949). Delay networks having maximally flat frequency characteristics. Proceedings of the IEE-Part III: Radio and Communication Engineering *96*, 487–490.

Table S1: Summary of disease labels across downstream evaluation datasets. Only test set counts are shown for HSP. Positive and negative sample counts are shown for each dataset.

Disease	**HSP** Positive / Negative	**SHHS** Positive / Negative	**WSC** Positive / Negative	**SSPH** Positive / Negative	**SRRSH** Positive / Negative
Arthritis	141 / 612	–	893 / 1677	–	–
Atrial fibrillation/flutter	53 / 417	–	–	–	–
Bronchitis	116 / 1074	437 / 7861	–	–	–
Cerebrovascular disease	101 / 1427	198 / 5455	58 / 2512	–	–
COPD	48 / 1431	258 / 8133	54 / 2516	–	–
Coronary heart disease	211 / 1315	660 / 5106	246 / 2324	23 / 966	–
Depression	328 / 994	–	–	316 / 673	–
Diabetes	248 / 1460	404 / 5115	342 / 2228	336 / 4970	86 / 903
GERD	122 / 114	–	–	–	–
Heart failure	35 / 1129	98 / 5421	29 / 2541	–	–
High cholesterol	128 / 203	–	–	–	–
Hypertension	772 / 931	4261 / 4179	1000 / 1569	296 / 5022	213 / 776
PTSD	52 / 772	–	–	–	–
Urologic or kidney problem	31 / 963	–	–	–	–

Table S2: Diagnostic performance (AUROC) of SleepFounder across diseases and cohorts, stratified by physiological systems. Values represent AUROC from the fine-tuned model.

System	**Disease**	**Internal**	**HSP**	**SHHS**	**WSC**	**SSPH**	**SRRSH**
Digestive system	Gastroesophageal reflux disease	0.8861	–	–	–	–	–
Cardiovascular system	Heart failure	0.8811	0.8514	0.8147	0.8033	–	–
	Coronary heart disease	0.8144	0.7848	0.7176	0.7598	–	0.7493
	Cerebrovascular disease	0.7755	0.6884	0.6898	0.6613	–	–
	Atrial fibrillation/flutter	0.7682	0.7671	–	–	–	–
	Hypertension	0.7213	0.7195	0.7007	0.7068	0.6724	0.7330
Endocrine and metabolic	High cholesterol	0.8304	–	–	–	–	–
	Diabetes	0.7845	0.7462	0.7224	0.7647	0.7308	0.7410
Musculoskeletal	Chronic pain	0.8102	0.7067	–	–	–	–
	Arthritis	0.7359	0.7253	–	0.6797	–	–
Respiratory system	Chronic obstructive pulmonary disease	0.7443	0.7708	0.7215	0.7401	–	–
Mental and behavioral disorders	Bipolar disorder	0.8110	0.7156	–	–	–	–
	Depression	0.7274	0.7313	–	–	–	–
	Post-traumatic stress disorder	0.7236	0.7396	–	–	–	–
	Anxiety	0.7168	0.6985	–	–	–	–
Genitourinary	Urologic or kidney problem	0.7365	0.7186	–	–	–	–

Appendix A Disease prevalence and diagnostic results

Table S1 summarizes disease prevalence across all cohorts. The corresponding diagnostic performance of SleepFounder for each condition is presented in Table S2.

Appendix B Data preprocessing

Heartbeat Component To capture heartbeat dynamics, we adopted inter-beat intervals (IBIs) as part of the model input. This choice allows us to preserve essential temporal information while significantly reducing the input signal's sampling frequency. For ECG signals, we first denoised the data and detected R peaks using NeuroKit2[42], from which the RR intervals were extracted as the IBIs. For BCG signals, we detected J peaks using the method proposed in[43], and the resulting JJ intervals were used as the IBIs. Once the raw IBIs were obtained, we applied the `get_nn_intervals` function from HRVAnalysis[44] to suppress outliers and artifacts, mitigating their

impact on the model. Finally, we linearly interpolated the cleaned IBI sequence at a fixed sampling rate of 4 Hz to generate a continuous heartbeat signal.

Respiratory Component The respiratory signal utilized in this work was extracted from the abdominal respiratory effort (ABD effort) in PSG recordings and from raw BCG signals. To ensure consistency and quality across modalities, a four-step preprocessing pipeline was employed, consisting of bandpass filtering, temporal resampling, integration (if necessary), and normalization. First, a third-order Bessel bandpass filter[45] with a passband of $\frac{1}{10}$Hz to $\frac{1}{3}$Hz was applied to suppress high-frequency noise and low-frequency drift. Next, all respiratory signals were resampled to 4 Hz to ensure alignment in temporal granularity. Due to the nature of BCG signals capturing derivative-like respiratory motion, a temporal integration step was applied to approximate respiratory volume and reduce phase mismatch relative to the ABD effort signal. This adjustment is not required for ABD-derived signals, which already represent respiratory amplitude. In the final step, all processed respiratory signals were standardized using z-score normalization to address inter-sensor variability and unify amplitude scales across modalities.

Electroencephalography spectrogram reconstruction The selection of EEG channels was prioritized based on their sensitivity to key sleep-related features such as sleep spindles, slow waves, and rapid eye movement activity. Specifically, we selected channels in the following order: $C3/C4 > F3/F4 > O1/O2$. The EEG signals corresponding to these channels were preprocessed into spectrograms using the Python LSPOpt package [5], which provides an optimal Wigner spectrum estimate for locally stationary processes (LSPs). To ensure consistent feature scaling, we applied log transformation and normalization. Specifically, the spectrogram was normalized by the standard deviation and mean of the spectrogram values across frequency bands and time segments. Additionally, linear interpolation was applied to adjust the frequency resolution between 0.3 Hz and 35 Hz. The entire preprocessing process is described by Equation (1).

$$S = \frac{10\log_{10}(\text{LSPOpt}(x) + 1e^{-13}) - \mu}{\sigma}, \tag{1}$$

where x represents the raw EEG signal, LSPOpt(x) denotes the locally stationary process estimation of x using the Wigner spectrum, μ is the mean, and σ is the standard deviation of the spectrogram values across frequency bands and time segments. The spectrogram values were then linearly interpolated between the frequencies 0.3 Hz and 35 Hz, and the logarithmic transformation $10\log_{10}$ was applied to scale the spectrogram values.

Peripheral capillary oxygen desaturation prediction SpO_2 represents the percentage of oxygen-saturated hemoglobin in the blood and reflects the efficiency of oxygen transport from the lungs to peripheral tissues, which is essential for maintaining normal physiological function. After resampling the SpO2 signal to a 1 Hz sampling frequency, the validity of each signal point was assessed by comparing it with its neighboring points. Points exhibiting a drop rate greater than 5 or a recovery rate greater than 10 were considered unreliable and replaced with a NaN value. If more than 40% of the data points within a given night were deemed erroneous, the entire SpO2 signal for that night was discarded. Instead of using the raw SpO2 signal, we computed oxygen desaturation as the learning target. Oxygen desaturation refers to the decrease in oxygen saturation relative to the highest local peak, capturing episodes of hypoxemia. This approach mitigates the effects of individual baseline differences in SpO2 levels, offering a more

[5] https://github.com/hbldh/lspopt

generalized representation of oxygen fluctuations. The local maximum used for this calculation was determined within a window of 100 seconds.