Supplementary Information for

A Unified Flexible Large Polysomnography Model for Sleep Staging and Mental Disorder Diagnosis

Guifeng Deng[1,2], Mengfan Niu[1], Shuying Rao[1,2], Yuxi Luo[3], Jianjia Zhang[3], Junyi Xie[1], Zhenghe Yu[1], Wenjuan Liu[1], Junhang Zhang[1], Sha Zhao[4], Gang Pan[4], Xiaojing Li[1,4], Wei Deng[1,4], Wanjun Guo[1,4], Yaoyun Zhang[5], Tao Li[1,4*], Haiteng Jiang[1,4,6*]

[1]Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, School of Brain Science and Brain Medicine, and Liangzhu Laboratory, Zhejiang University School of Medicine, Hangzhou, 310058, China.

[2]College of Biomedical Engineering & Instrument Science, Zhejiang University, Hangzhou, 310058, China.

[3]School of Biomedical Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, 518100, China.

[4]MOE Frontier Science Center for Brain Science and Brain-machine Integration, State Key Lab of Brain-Machine Intelligence, Zhejiang University, Hangzhou 311121, China.

[5] School of Biomedical Informatics, University of Texas at Dallas, 800 W Campbell Rd, Richardson, TX 75080, USA.

[6] Department of Psychiatry and Mental Health, Wenzhou Medical University, Wenzhou 325035, Zhejiang Province, China.

*Corresponding authors:

Tao Li Email: litaozjusc@zju.edu.cn

Haiteng Jiang Email: h.jiang@zju.edu.cn

Supplementary Methods

Supplementary Figures

Supplementary Tables

Supplementary Notes

Supplementary Methods

Datasets

To the best of our knowledge, this work represents the most extensive use of data for training a large-scale sleep staging model to date. We aggregated approximately 24,000 full-night PSG recordings, encompassing roughly 220,500 hours of sleep data, from 16 publicly available datasets as source domains for training. A summary of the public datasets is provided below.

(1) The Apnea Positive Pressure Long-term Efficacy Study (APPLES) is a multi-center dataset that includes overnight PSG recordings from 1,104 patients with obstructive sleep apnea syndrome (OSAS). The dataset includes four EEG and two EOG channels, and was scored using Rechtschaffen and Kales (R&K) criteria.

(2) The Danish Center for Sleep Medicine (DCSM) dataset consists of 255 randomly selected and fully anonymized overnight lab-based PSG recordings from patients visiting the DCSM for the diagnosis of non-specific sleep related disorders. The dataset includes six EEG and two EOG channels, and was scored according to the AASM criteria.

(3) The Dreem Open Dataset (DOD) consists of two subsets, DOD-H and DOD-O. DOD-H comes from French Armed Forces Biomedical Research Institute's (IRBA), and contains PSG recordings from 25 healthy volunteers. DOD-O comes from the Stanford Sleep Medicine Center, and contains PSG recordings from 56 OSAS patients. Both datasets contain twelve and eight EEG channels, respectively. For experimentation, we specifically selected the three and five channels that overlap with the AASM criteria. Both contain two EOG channels and were scored according to the AASM criteria.

(4) Haaglanden Medisch Centrum (HMC) dataset consists of 151 randomly selected whole-night PSG of different sleep disorders. The dataset includes four EEG and two EOG channels, and was scored according to AASM criteria.

(5) The Institute of Systems and Robotics, University of Coimbra (ISRUC) dataset consists of 126 PSG recordings from the Sleep Medicine Center of the Hospital of the University of Coimbra, Portugal. The dataset comprises three groups of data. Data in group one concerning 100 subjects, with one recording session per subject. Data in group two is gathered from 8 subjects and two recording sessions were performed per subject. Data in group three is collected from one recording session related to 10 healthy subjects. The dataset includes six EEG and two EOG channels and was scored according to the AASM criteria.

(6) The St. Vincent's University Hospital (SVUH) dataset contains 25 full overnight PSG with suspected sleep-disordered breathing. The dataset contains two EEG and two EOG channels and was scored according to R&K criteria.

(7) P2018 (You Snooze You Win: The PhysioNet/Computing in Cardiology Challenge 2018) dataset was contributed by the Massachusetts General Hospital's (MGH) Computational Clinical Neurophysiology Laboratory (CCNL),

and the Clinical Data Animation Laboratory (CDAC). The dataset consists of 994 training examples and 989 test examples, with only the training data having labels publicly available. We only use the training data, which includes 6 EEG and 1 EOG channels, and uses the AASM criteria for sleep staging.

(8) The Stanford Technology Analytics and Genomics in Sleep (STAGES) dataset was collected on 1500 patients evaluated for sleep disorders from six centers. The dataset contains six EEG and two EOG channels and is annotated based on AASM criteria.

(9) The Apnea, Bariatric surgery, and CPAP (ABC) dataset includes 80 patients with severe OSAS, with six EEG and two EOG channels, being scored based on AASM criteria.

(10) The Nationwide Children's Hospital Sleep DataBank (NCHSDB) has 3,984 pediatric sleep studies on 3,673 unique patients conducted at NCH in Columbus, Ohio, USA between 2017 and 2019. The dataset includes six EEG and two EOG channels and using AASM criteria.

(11) The Home Positive Airway Pressure (HOMEPAP) study was a multi-center dataset that enrolled 373 patients with suspected moderate and severe OSAS. Subjects were randomized to lab-based and home-based management. We only use the lab-based subset as it includes the channels we need. The lab-based subset includes six EEG and two EOG channels with AASM criteria annotation.

(12) The Childhood Adenotonsillectomy Trial (CHAT) is a multi-center dataset that enrolled 1447 children with mild to moderate OSAS. The dataset includes six EEG and two EOG channels with AASM criteria annotation.

(13) The Cleveland Children's Sleep and Health Study (CCSHS) dataset consists of 515 PSG recordings from 907 children aged between 8 and 11 years old. The dataset includes two EEG and two EOG channels, and was scored according to AASM criteria.

(14) The Cleveland Family Study (CFS) is a family-based study of sleep apnea worldwide, compring 730 overnight PSG from 2284 indivisuals. The dataset includes two EEG and two EOG channels with AASM criteria.

(15) The MROS dataset enrolled 5994 men 65 years or older at six clinical centers. The dataset consists of 3929 PSG recordings with two EEG and two EOG channals, and was scored according to AASM criteria.

(16) The Sleep Heart Health Study (SHHS) is a multi-center cohort study implemented by the National Heart Lung & Blood Institute, consisting of two subset. SHHS-1 and SHHS-2 contain 5793 and 2651 overnight PSG, respectively. The dataset contains two EEG and EOG channels and was scored according to R&K criteria.

We test the model on two private datasets from different clinical centers to evaluate the cross-center generalization performance of our methods. The discription of each dataset is provided below.

(1) The HANG7 dataset was acquired from Affiliated Mental Health Center & Hangzhou Seventh People's Hospital, Zhejiang University School of Medicine.

Data collection was conducted at Zhejiang University with Institutional Review Board approval, and written consent was obtained from all subjects or their caregivers. The dataset comprises PSG recordings from 127 subjects, including 33 healthy controls, 51 patients diagnosed with narcolepsy, and 43 patients exhibiting hypersomnia symptoms associated with anxiety and depression but not meeting the diagnostic criteria for narcolepsy. The dataset includes six EEG and two EOG channels sampled at a frequency of 512 Hz. PSG recordings were scored by experienced clinicians according to AASM criteria.

(2) The SYSU dataset includes two groups: 20 healthy controls and 24 patients with major depressive disorder (MDD). All subjects were equipped with the same device to collect overnight PSG signals. This study received approval from the Ethics Committee of Guangdong 999 Brain Hospital (approval number: 2020-010-059). The experiments involving healthy individuals were conducted at the sleep laboratory of Sun Yat-sen University, encompassing 80 PSG recordings from 20 subjects sampled at 500 Hz. The MDD dataset comprised 24 PSG recordings from 24 MDD, who were diagnosed by two experienced psychiatrists based on the criteria of Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV). The scores of 24-item Hamilton Depression Scale (HAMD-24) and the self-rating Depression Scale (SDS) were 22.6±6.20 and 65.57±9.53, respectively. All patients were without drug abuse, suicide risk, pregnancy, present or history of head injuries, seizures, or epilepsy. The EEG signals were sampled at 256 Hz. Both groups included six EEG and two EOG channels and were scored according to the AASM scoring manual by two well-trained sleep technologists.

Preprocessing

For all PSG datasets, we selected eight EEG/EOG channels recommended by the AASM criteria for sleep staging. No other channels or data types present in the datasets were utilized. The chosen eight channels included F3-M2, F4-M1, C3-M2, C4-M1, O1-M2, O2-M1, E1-M2, and E2-M1.

All the PSG recordings were initially subjected to a fourth-order bandpass filter (0.3 Hz to 35 Hz) and subsequently resampled at a rate of 100 Hz. Finally, Z-score normalization was applied individually to each channel of every PSG recording:

$$x[c] = \frac{x[c] - \mathrm{mean}(x[c])}{\mathrm{std}(x[c])}, c \in C \quad (1)$$

where x represents a single PSG recording, C denotes the set of channels for that recording. The samples were clamped to the range $[-10,10]$ after Z-score normalization to minimize the impact of outliers.

We adhered to the current AASM sleep staging standards. For datasets originally labeled according to the R&K standards, we followed the conventional approach of merging stages N3 and N4 into a single N3 stage. Additionally, we removed any sleep epochs without labels, which typically indicated sensor detachment, sleep interruptions, or other anomalies. After removing such segments, the data was divided into two distinct segments at that specific point. Table 2 shows the class distribution across the

datasets.

Classification Metrics with Probabilistic Extensions

Let n be the total number of 30-s epochs retained, C be the number of sleep-stage classes. For each epoch $i = 1, \dots, n$

$$y_i^{\text{true}} = \left(y_{i,1}^{\text{true}}, \dots, y_{i,C}^{\text{true}}\right) \text{ and } y_i^{\text{pred}} = \left(y_{i,1}^{\text{pred}}, \dots, y_{i,C}^{\text{pred}}\right)$$

are the true-label and predicted-probability vectors (each in $[0,1]^C$ and summing to 1).

Probabilistic Accuracy

The probabilistic accuracy is an extended metric that calculates the average match between the true and predicted probabilistic distributions across all samples. It quantifies how well the predicted probabilities align with the true probabilities on a per-sample basis.

$$\text{Acc} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{C} y_{i,k}^{\text{true}} y_{i,k}^{\text{pred}} \times 100\% \tag{2}$$

This metric essentially computes the weighted sum of the element-wise product of the true and predicted probability distributions for each sample, then averages these values across all samples.

Probabilistic Cohen's Kappa

The probabilistic Cohen's Kappa extends the traditional Cohen's Kappa to accommodate probabilistic predictions. It measures the agreement between the true and predicted distributions, accounting for the possibility of multiple correct classes per sample.

First construct the soft confusion matrix

$$M_{ab} = \sum_{i=1}^{n} y_{i,a}^{\text{true}}\, y_{i,b}^{\text{pred}} \quad (a, b = 1, \dots, C) \tag{3}$$

and denote its total weight $N = \sum_{a,b} M_{ab}$. Then the observed agreement is

$$P_o = \frac{1}{N}\sum_{k=1}^{C} M_{kk} \tag{4}$$

and the chance agreement

$$P_e = \frac{1}{N^2}\sum_{k=1}^{C}\left(\sum_{b=1}^{C} M_{kb}\right)\left(\sum_{a=1}^{C} Mak\right) \tag{5}$$

Finally,

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{6}$$

This extension allows for a more nuanced assessment of model performance when dealing with probabilistic predictions and scenarios where multiple classes may be partially correct for a given sample.

It should be noted that when the probability distributions are deterministic (i.e., the

probability is concentrated in a single class for each sample, such as [1,0,0], [0,1,0], or [0,0,1]), the probabilistic accuracy and Cohen’s Kappa reduce to their traditional counterparts. In such cases, the probabilistic metrics yield identical results to the conventional accuracy and Cohen's Kappa, ensuring consistency with standard evaluation practices when dealing with hard classifications.

Supplementary Figures

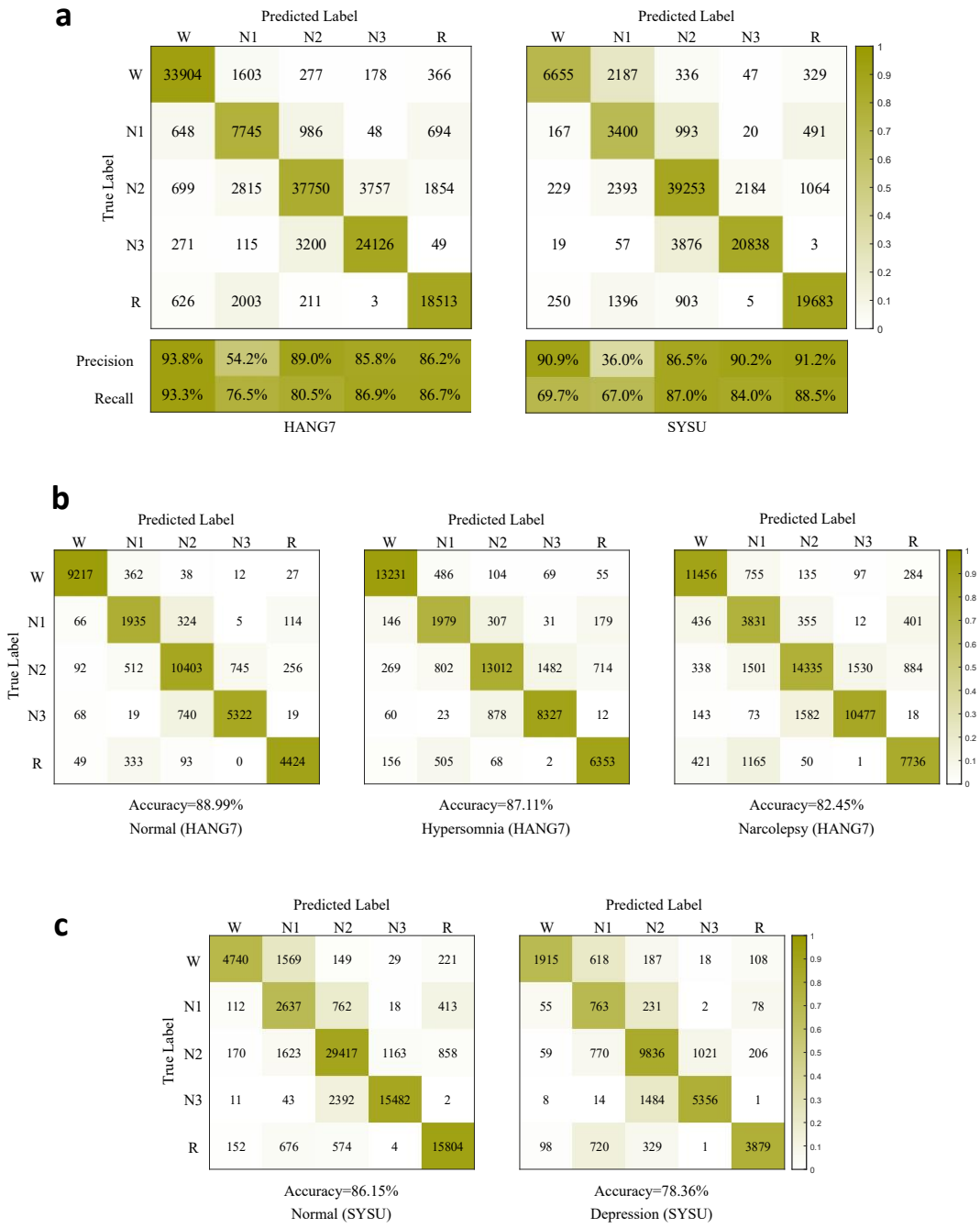

Figure S1: Cross-center sleep staging performance of LPSGM. Panel (a) presents the confusion matrices of LPSGM on the HANG7 and SYSU datasets. The values in each cell represent the number of 30-second epochs classified into each sleep stage, with darker shades indicating higher counts. Precision and recall scores for each sleep stage are reported below. Panel (b) shows confusion matrices for different subject groups within the HANG7 dataset. Panel (c) displays confusion matrices for different subject groups within the SYSU dataset.

Supplementary Tables

Table S1: Definitions of hypnogram metrics.

Hypnogram Metrics		Definition
Sleep Latency (min)		
N1		Default to 0.
N2		The time from the onset of the first Stage N1 sleep to the onset of the first Stage N2 sleep.
N3		The time from the onset of the first Stage N1 sleep to the onset of the first Stage N3 sleep.
REM		The time from the onset of the first Stage N1 sleep to the onset of the first REM sleep episode.
Sleep Duration (min)		
TST (Total Sleep Time)		The total duration of sleep during the sleep period, calculated as the sum of all NREM and REM sleep stages.
REM		The cumulative duration of all REM sleep episodes during the sleep period.
NREM		The cumulative duration of all non-REM sleep stages (N1, N2, and N3) during the sleep period.
SWS		The cumulative duration of Stage N3 sleep (slow-wave sleep) during the sleep period.
Sleep Stages		
W (SPT)	Episodes (#)	The number of wake episodes (Wake after Sleep Onset, WASO) during the sleep period time (SPT).
	Duration (min)	The total time spent awake during the sleep period time (SPT).
R	Duration (min)	Identical to NREM.
	TST (%)	The proportion of REM sleep relative to the total sleep time (TST).
N1	Duration (min)	The cumulative duration of all Stage N1 sleep episodes during the sleep period.
	TST (%)	The proportion of Stage N1 sleep relative to the total sleep time (TST).
N2	Duration (min)	The cumulative duration of all Stage N2 sleep episodes during the sleep period.
	TST (%)	The proportion of Stage N2 sleep relative to the total sleep time (TST).
N3	Duration (min)	The cumulative duration of all Stage N3 sleep episodes during the sleep period.
	TST (%)	The proportion of Stage N3 sleep (SWS) relative to the total sleep time (TST).

Table S2: Overview of datasets used in our experiments.

Datasets		Recordings	Annotation	EEG						EOG		Channels	Sample Rate	
				F3-M2	F4-M1	C3-M2	C4-M1	O1-M2	O2-M1	E1-M2	E2-M1		EEG	EOG
APPLES		1067	R&K			√	√	√	√	√	√	6	100	100
DCSM		255	AASM	√	√	√	√	√	√	√	√	8	256	256
DOD	DOD-H	25	AASM	√	√	√				√	√	5	250	250
	DOD-O	56	AASM	√		√	√	√	√	√	√	7	250	250
HMC		151	AASM		√	√	√		√	√	√	6	256	256
ISRUC		126	AASM	√	√	√	√	√	√	√	√	8	200	200
SVUH		25	R&K			√	√			√	√	4	128	64
P2018		994	AASM	√	√	√	√	√	√	√		7	200	200
STAGES	BOGN	85	AASM	√	√	√	√	√	√	√	√	8	200	200
	STNF	525		√	√	√	√	√	√	√	√	8	256	256
	GSDV	288		√	√	√	√	√	√	√	√	8	200	200
	MSTR	286		√	√	√	√	√	√	√	√	8	256	256
	GSBB	38		√	√	√	√	√	√	√	√	8	200	200
	GSLH	51		√	√	√	√	√	√	√	√	8	200	200
	GSSA	34		√	√	√	√	√	√	√	√	8	200	200
	GSSW	131		√	√	√	√	√	√	√	√	8	200	200
	MSMI	61		√	√	√	√	√	√	√	√	8	200	200
	MSNF	35		√	√	√	√	√	√	√	√	8	200	200
	MSQW	145		√	√	√	√	√	√	√	√	8	200	200
	MSTH	31		√	√	√	√	√	√	√	√	8	256	256
	STLK	156		√	√	√	√	√	√	√	√	8	500	500
ABC		132	AASM	√	√	√	√	√	√	√	√	8	256	256

NCHSDB		3947	AASM	√	√	√	√	√	√	√	√	8	256/400/512	256/400/512
HOMEPAP		245	AASM	√	√	√	√	√	√	√	√	8	200/256	200/256
CHAT		1638	AASM	√	√	√	√	√	√	√	√	8	200	200
CCSHS		515	AASM			√	√			√	√	4	128	128
CFS		730	AASM			√	√			√	√	4	128	128
MROS		3929	AASM			√	√			√	√	4	256	256
SHHS	SHHS-1	5793	R&K			√	√			√	√	4	125	50
	SHHS-2	2651				√	√			√	√	4	128	32
HANG7		127	AASM	√	√	√	√	√	√	√	√	8	512	512
SYSU		104	AASM	√	√	√	√	√	√	√	√	8	500/256	500/256

Table S3: Number of sleep stages of the datasets after preprocessing.

Datasets		Total	Epochs					Ratio %				
			W	N1	N2	N3	R	W	N1	N2	N3	R
APPLES		1049110	256128	147217	481003	24295	140467	24.4	14	45.8	2.3	13.4
DCSM		304266	79636	21140	113027	43637	46826	26.2	6.9	37.1	14.3	15.4
DOD	DOD-H	24662	3037	1505	11879	3514	4727	12.3	6.1	48.2	14.2	19.2
	DOD-O	54197	10660	2898	26650	5683	8306	19.7	5.3	49.2	10.5	15.3
HMC		137243	23686	15548	50083	26671	21255	17.3	11.3	36.5	19.4	15.5
ISRUC		107784	23198	14254	34661	21489	14182	21.5	13.2	32.2	19.9	13.2
SVUH		20774	4707	3016	3403	7658	1990	22.7	14.5	16.4	36.9	9.5
P2018		892262	157945	136978	377870	102592	116877	17.7	15.4	42.3	11.5	13.1
STAGES	BOGN	75930	23185	4623	29819	8236	10067	30.5	6.1	39.3	10.8	13.3
	STNF	592027	202352	57652	171620	52147	108256	34.2	9.7	29	8.8	18.3
	GSDV	218608	54699	13363	114794	10194	25558	25	6.1	52.5	4.7	11.7
	MSTR	221466	42522	23789	103118	24201	27836	19.2	10.7	46.6	10.9	12.6
	GSBB	29843	7954	2469	14423	1668	3329	26.7	8.3	48.3	5.6	11.2
	GSLH	32118	8540	2684	16485	1597	2812	26.6	8.4	51.3	5	8.8
	GSSA	27751	6998	920	15233	881	3719	25.2	3.3	54.9	3.2	13.4
	GSSW	90341	25790	5944	44213	3549	10845	28.5	6.6	48.9	3.9	12
	MSMI	45892	8035	4227	22959	4847	5824	17.5	9.2	50	10.6	12.7
	MSNF	27061	4905	1385	13454	4012	3305	18.1	5.1	49.7	14.8	12.2
	MSQW	113085	25039	13622	53319	8125	12980	22.1	12	47.1	7.2	11.5
	MSTH	23682	5036	2146	12033	1755	2712	21.3	9.1	50.8	7.4	11.5
	STLK	154691	29429	11937	76878	11426	25021	19	7.7	49.7	7.4	16.2
ABC		133000	30938	19296	52334	11761	18671	23.3	14.5	39.3	8.8	14

NCHSDB		3661376	665063	128183	1382551	874762	610817	18.2	3.5	37.8	23.9	16.7
HOMEPAP		229604	63395	24759	86718	22645	32087	27.6	10.8	37.8	9.9	14
CHAT		1957293	469804	119436	628932	464993	274128	24	6.1	32.1	23.8	14
CCSHS		691401	212027	19221	249698	110191	100264	30.7	2.8	36.1	15.9	14.5
CFS		866204	321333	26394	306264	111937	100276	37.1	3	35.4	12.9	11.6
MROS		5373725	2609224	218233	1735010	278620	532638	48.6	4.1	32.3	5.2	9.9
SHHS	SHHS-1	5863207	1691288	217583	2397460	739403	817473	28.8	3.7	40.9	12.6	13.9
	SHHS-2	3192507	1208326	111456	1147780	313790	411155	13.9	3.5	36	9.8	12.9
HANG7		142450	36337	10121	46875	27761	21356	25.5	7.1	32.9	19.5	15
SYSU		106778	9554	5071	45123	24793	22237	8.9	4.7	42.3	23.2	20.8

Supplementary Notes

Acknowledgements

The Apnea Positive Pressure Long-term Efficacy Study (APPLES) was supported by the National Heart, Lung, and Blood Institute (U01HL68060).

This research has been conducted using the STAGES - Stanford Technology, Analytics and Genomics in Sleep Resource funded by the Klarman Family Foundation. The investigators of the STAGES study contributed to the design and implementation of the STAGES cohort and/or provided data and/or collected biospecimens, but did not necessarily participate in the analysis or writing of this report. The full list of STAGES investigators can be found at the project website.

The Apnea, Bariatric surgery, and CPAP study (ABC Study) was supported by National Institutes of Health grants R01HL106410 and K24HL127307. Philips Respironics donated the CPAP machines and supplies used in the perioperative period for patients undergoing bariatric surgery.

The Home Positive Airway Pressure study (HomePAP) was supported by the American Sleep Medicine Foundation 38-PM-07 Grant: Portable Monitoring for the Diagnosis and Management of OSA.

The Sleep Heart Health Study (SHHS) was supported by National Heart, Lung, and Blood Institute cooperative agreements U01HL53916 (University of California, Davis), U01HL53931 (New York University), U01HL53934 (University of Minnesota), U01HL53937 and U01HL64360 (Johns Hopkins University), U01HL53938 (University of Arizona), U01HL53940 (University of Washington), U01HL53941 (Boston University), and U01HL63463 (Case Western Reserve University).

The Pediatric Adenotonsillectomy Trial for Snoring (PATS) study was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (1U01HL125307, 1U01HL125295).

The Childhood Adenotonsillectomy Trial (CHAT) was supported by the National Institutes of Health (HL083075, HL083129, UL1-RR-024134, UL1 RR024989).

The Cleveland Children's Sleep and Health Study (CCSHS) was supported by grants from the National Institutes of Health (RO1HL60957, K23 HL04426, RO1 NR02707, M01 Rrmpd0380-39).

The Cleveland Family Study (CFS) was supported by grants from the National Institutes of Health (HL46380, M01 RR00080-39, T32-HL07567, RO1-46380).

The Mignot Nature Communications research was mostly supported by a grant from Jazz Pharmaceuticals to E.M. Additional funding came from: NIH grant R01HL62252 (to P.E.P.); Ministry of Science and Technology 2015CB856405 and National Foundation of Science of China 81420108002,81670087 (to F.H.); H. Lundbeck A/S, Lundbeck Foundation, Technical University of Denmark and Center for Healthy Aging, University of Copenhagen (to P.J. and H.B.D.S). Additional support was provided by the Klarman Family, Otto Mønsted, Stibo, Vera & Carl Johan Michaelsens, Knud Højgaards, Reinholdt W. Jorck and Hustrus and Augustinus Foundations (to A.N.O.).

NCH Sleep DataBank was supported by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under Award Number R01EB025018.

The National Heart, Lung, and Blood Institute provided funding for the ancillary MrOS Sleep Study, "Outcomes of Sleep Disorders in Older Men," under the following grant numbers: R01 HL071194, R01 HL070848, R01 HL070847, R01 HL070842, R01 HL070841, R01 HL070837, R01 HL070838, and R01 HL070839.

The National Sleep Research Resource was supported by the U.S. National Institutes of Health, National Heart Lung and Blood Institute (R24 HL114473, 75N92019R002).