SLEEPJ, 2020, 1–12

doi: 10.1093/sleep/zsz306
Advance Access Publication Date: 21 December 2019
Original Article

ORIGINAL ARTICLE

Sleep staging from electrocardiography and respiration with deep learning

Haoqi Sun[1,], Wolfgang Ganglberger[1], Ezhil Panneerselvam[1], Michael J. Leone[1], Syed A. Quadri[1,], Balaji Goparaju[1], Ryan A. Tesh[1], Oluwaseun Akeju[2], Robert J. Thomas[3] and M. Brandon Westover[1,*]

[1]Department of Neurology, Massachusetts General Hospital, Boston, MA, [2]Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital, Boston, MA and [3]Division of Pulmonary, Critical Care & Sleep, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA

*Corresponding author. M. Brandon Westover, Department of Neurology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114. Email: mwestover@mgh.harvard.edu.

Abstract

Study Objectives: Sleep is reflected not only in the electroencephalogram but also in heart rhythms and breathing patterns. We hypothesized that it is possible to accurately stage sleep based on the electrocardiogram (ECG) and respiratory signals.

Methods: Using a dataset including 8682 polysomnograms, we develop deep neural networks to stage sleep from ECG and respiratory signals. Five deep neural networks consisting of convolutional networks and long- and short-term memory networks are trained to stage sleep using heart and breathing, including the timing of R peaks from ECG, abdominal and chest respiratory effort, and the combinations of these signals.

Results: ECG in combination with the abdominal respiratory effort achieved the best performance for staging all five sleep stages with a Cohen's kappa of 0.585 (95% confidence interval ±0.017); and 0.760 (±0.019) for discriminating awake vs. rapid eye movement vs. nonrapid eye movement sleep. Performance is better for younger ages, whereas it is robust for body mass index, apnea severity, and commonly used outpatient medications.

Conclusions: Our results validate that ECG and respiratory effort provide substantial information about sleep stages in a large heterogeneous population. This opens new possibilities in sleep research and applications where electroencephalography is not readily available or may be infeasible.

Statement of Significance

Large datasets of sleep recordings (PSGs) have the potential to provide robust information for sleep research. Sleep staging is traditionally done mostly using electroencephalograms. However, the interaction of the brain and the body is also an important factor during sleep. How accurately sleep can be staged from heart beats and respiration remains unknown. Existing work is limited to a small numbers of PSGs, limiting the validity and robustness of conclusions. Here, we developed models for sleep staging using electrocardiogram and respiratory effort signals based on a large sleep dataset, containing 8682 PSGs. The models open new possibilities in translational research where electroencephalography is not available, such as in critical care units.

Key words: electrocardiography; respiration; sleep stages; deep learning

Submitted: 9 July, 2019; **Revised:** 13 November, 2019

Introduction

Characterizing sleep has primarily relied on analysis of the electroencephalogram (EEG), supplemented by the electrooculogram and chin electromyogram [1]. Three distinct states are readily discernable through such analysis: wake, rapid eye movement (REM) sleep and non-REM (NREM) sleep [1]. Three stages of progressive depth (N1, N2, and N3) are conventionally differentiated in NREM [1]. EEG delta wave oscillations (about 1–4 Hz) dominate deeper NREM sleep, while sleep spindles (about 10–15 Hz) and theta wave oscillations (about 4–8 Hz) dominate lighter NREM sleep [1].

Cortical, subcortical, and brainstem systems are highly interactive throughout sleep [2–5], and their activity couples autonomic activity with cortical activity measurable by EEG. Examples include strong sinus arrhythmia [3], blood pressure dipping [6], and stable breathing or stable flow-limitation [7]. REM sleep is characterized by highly recognizable respiratory rate and tidal volume fluctuations, and by surges in heart rate and blood pressure [8]. Wake demonstrates dominance of low-frequency heart rate variability and large amplitude movements. These observations suggest that accurate sleep staging might be possible from non-EEG signals influenced by the autonomic nervous system, such as the electrocardiogram (ECG) or respiratory signals.

An accurate non-EEG method for sleep state characterization would have several advantages. For example, the ECG is recorded continuously in numerous medical settings, especially in hospitalized patients. Wearable devices increasingly measure ECG and respiration [9–11]. Cardiorespiratory signals may be obtainable in a number of ways, including contact recordings such as Withings, ballistocardiogram [12], or non-contact radar-type applications such as EarlySense [13], and SleepScore [14], etc. On the other hand, EEG can be highly abnormal in medically ill populations, making standard analysis difficult [15].

We hypothesize that deep learning approaches can be used to accurately estimate sleep states during ECG and respiration. We previously showed that deep neural networks can learn to score conventional sleep stages based on EEG signals obtained during overnight PSG with Cohen's kappa of 0.805, comparable to the agreement between human sleep scoring experts [16]. Here, our approach is based on a convolutional neural network (CNN) in combination with long–short-term memory (LSTM) recurrent neural network. It is trained on a large clinical dataset, which also accounts for patient heterogeneity, spanning a wide range of ages, apnea severities, medications, and sleep disorders.

Methods

Dataset

The Partners Institutional Review Board approved retrospective analysis of polysomnograms (PSG), acquired in the Sleep Laboratory at Massachusetts General Hospital from 2009 to 2016, without requiring additional consent for use in this study. PSGs were recorded adhering to American Academy of Sleep Medicine (AASM) standards. Each PSG includes one ECG channel and two respiratory effort channels recorded from chest and abdomen belts. The sampling frequency is 200 Hz for all signals. The dataset contains three major types of sleep tests: diagnostic, full-night continuous positive airway pressure (CPAP), and split-night CPAP. PSGs were annotated in 30-s nonoverlapping time windows according to AASM standards as one of the five stages: wake (W), non-REM stage 1 (N1), non-REM stage 2 (N2), non-REM stage 3 (N3), and rapid eye movement (REM). Seven sleep technicians in total annotated the dataset, with one technician per PSG. PSG was recorded using Grass Technologies equipment; Grass Technologies is now owned by Natus Neuro, CA, USA. The entire dataset includes 10,121 PSGs; 9644 were exported successfully without time mismatch or missing sleep stage annotations. We included atrial fibrillation cases because the deep learning network is intended to be used with heterogeneous data. We excluded PSGs with (1) fewer than 100 artifact-free 30-s time windows; (2) unreadable data; (3) missing the required signal channels; and (4) recordings with ≤2 sleep stages present whole night (usually due to sparse technician labeling), resulting in 8682 PSGs coming from 7208 unique patients. Thus, approximately 10% (962/9644) PSGs were excluded due to poor overall signal quality. These records were excluded at the outset, before training or testing the neural network, based on an automated assessment of signal quality and data completeness (see Methods: Preprocessing Section). The dataset is summarized in Table 1.

Preprocessing

Sleep staging was done in 30-s time windows following AASM standards. However, changes in heart rhythms and respiration often occur over longer time scales. For this reason, and to provide contextual information, our deep neural networks used information extending 120 s on both sides of each 30-s time window, creating a 270-s time window (4.5 min, nine 30-s time windows) centered on each 30-s time window to be scored. The goal of the deep neural networks presented herein is to classify the sleep stage of the middle 30-s time window using information from surrounding context. This is illustrated in Supplementary Figure S1 in the supplementary material. To clarify the terminology, note that we use "time windows" for the consecutive intervals on the signal that are assigned stages by sleep technicians and by our algorithm (normally termed "epochs" in the sleep literature). We will use the word "epoch" according to the deep learning literature to refer to a complete scan through the whole training dataset during model training. The combination of 270 s time windows and 30 s time steps means that our algorithm assigns a sleep stage to each 30 s time window, while also using 240 s of contextual information from the signal surrounding the central 30 s window, so that overall the window time is 270 s.

When using ECG as the input, we identified 270-s time windows with any voltage larger than 6 mV or with standard deviation of the entire 270 s smaller than 5 μV, as they are likely to represent nonphysiological artifacts. In each 270-s time window we extracted timings of R peaks [17] and converted the ECG to a binary sequence, where R peaks are indicated by 1 and all other points indicated by 0. The 270-s time windows with spurious R peaks were identified using the ADARRI [18] method. 270-s time windows with less than 20 R peaks/min were also identified, as they are not physiological. About 25% of the 270-s time windows were identified as artifact. In total, there were 5,964,359 270-s time windows.

When using chest and abdominal respiratory effort as the input, 270-s time windows with any voltage larger than 6 mV

Table 1. Dataset summary.

Characteristics	Value
Number of PSGs	8682
Number of patients	7208
Age: year, median (IQR)	53 (41–63)
Sex: number (percentage of all patients)	
Female	2997 (41.6%)
Male	4189 (58.1%)
Unknown due to human error	22 (0.3%)
BMI: kg/m^2, median (IQR)	31 (27–36)
Type of Test: number (percentage of all patients)	
Diagnostic	3571 (49.5%)
All night CPAP	1751 (24.3%)
PSG split night	1798 (24.9%)
Extended EEG-sleep montage	76 (1.1%)
Bedside	9 (0.1%)
Research	3 (0.04%)
Apnea-Hypopnea Index (AHI, events/hour): number (percentage of all patients)	
Normal (AHI < 5)	2879 (39.9%)
Mild (5 ≤ AHI < 15)	1995 (27.7%)
Moderate (15 ≤ AHI < 30)	1468 (20.4%)
Severe (AHI ≥ 30)	866 (12.0%)
Respiratory Disturbance Index (RDI, events/hour)	15.0 (5.8–28.4)
Periodic Limb Movement Index (PLMI, events/hour)	10.4 (3.1–28.3)
Outpatient medication listing, by category	
Systemic	4523 (62.7%)
Hypertension	2755 (38.2%)
Sleeping	2187 (30.3%)
Antidepressant	1874 (26.0%)
Neuroactive	1365 (18.9%)
Benzodiazepine	1297 (18.0%)
Diabetic	802 (11.1%)
RLS/PLMS	688 (9.5%)
Opiate	548 (7.6%)
Z-drug	348 (4.8%)
Stimulant	310 (4.3%)
Neuroleptic	282 (3.9%)
Herbal	280 (3.9%)

or standard deviation of the entire 270 s smaller than 10 μV were identified. Respiratory signals were down-sampled to 10 Hz. About 10% of all 270-s time windows were identified as artifact. In total, there were 6,847,246 270-s time windows for the chest signal; and 6,749,286 270-s time windows for the abdominal signal.

When using multiple signal modalities as the input, 270-s time windows where any signal modality meet the above criteria are identified as artifact.

Deep network architecture

We trained five deep neural networks based on the following input signals and their combinations: 1) ECG; 2) CHEST (chest respiratory effort); 3) ABD (abdominal respiratory effort); 4) ECG + CHEST; and 5) ECG + ABD. Each deep neural network contained a feed-forward CNN which learned features pertaining to each time window, and a recurrent neural network (RNN), in this case a LSTM network, to learn temporal patterns among consecutive time windows.

The CNN of the network is similar to that in Hannun et al. [19]. As shown in Figure 1A and B, the network for a single type of input signal, i.e. ECG, CHEST or ABD, consists of a convolutional layer, several residual blocks and a final output block. For a network with both ECG and CHEST/ABD as input signals (Figure 1C), we first fixed the weights of the layers up to the ninth residual block (gray) for the ECG network and similarly fixed up to the fifth residual block (gray) for the CHEST/ABD network, concatenated the outputs, and then fed this concatenation into a subnetwork containing five residual blocks and a final output block. The numbers of fixed layers were chosen so that the outputs of layers from different modalities have the same shape (after padding zeros), and were then concatenated.

The LSTM of the network has the same structure for different input signals. It is a bi-directional LSTM, where the context cells from the forward and backward directions are concatenated. For the network with ECG as input, the LSTM has two layers with 20 hidden nodes in each layer. For CHEST and ECG + CHEST, the LSTM has three layers with 100 hidden nodes in each layer. For ABD and ECG + ABD, the LSTM has two layers with 100 hidden nodes in each layer. The number of LSTM layers, number of hidden nodes, and dropout rate were determined by the method described in the next section.

Training and evaluating the network

To obtain an unbiased estimate of out-of-sample performance, we did five-fold cross validation. The 7208 unique patients were randomly split into five folds. We trained the model on four folds, and then tested the model on the left-out testing fold. Training and testing folds were constructed to always contain unique, nonoverlapping sets of patients. This procedure was repeated five times so that the five testing folds covered the whole dataset. The reported performance metrics are based on the pooled predictions across the five testing folds. For each split, we first train the CNN, and then train the LSTM using the outputs from the CNN. The objective function of both CNN and LSTM is cross-entropy, a measure of the distance between two categorical distributions for classification The LSTM is trained using sequences of 20 time windows (14 min). Note that the CNN is trained on time windows without artifacts, whereas the LSTM is trained on time windows including those with artifacts, so that the 20 time windows are consecutive, preserving the temporal context. We set the number of LSTM layers, number of hidden nodes, and the dropout rate as the combination that minimizes the objective function on the validation set. The networks were trained with a mini-batch size of 32, maximum number of epochs of 10, and learning rate 0.001 (as commonly used in deep learning). During training, we reduce the learning rate by 10% when the loss on the validation set does not decrease for three consecutive epochs. We stop training when the validation loss does not decrease for six consecutive epochs.

Some sleep stages occur more frequently than others. For example, people spend about 50% of sleep in N2 and 20% in N3. To prevent the network from simply learning to report the dominant stage, we weighed each 270-s input signal in the objective function by the inverse of the number of time windows in each sleep stage within the training set.

The reported performance metrics were all based on the pooled predictions from the five testing folds. We used Cohen's kappa, macro-F1 score, weighted macro-F1 score (weighted by the number of time windows in each sleep stage to account for

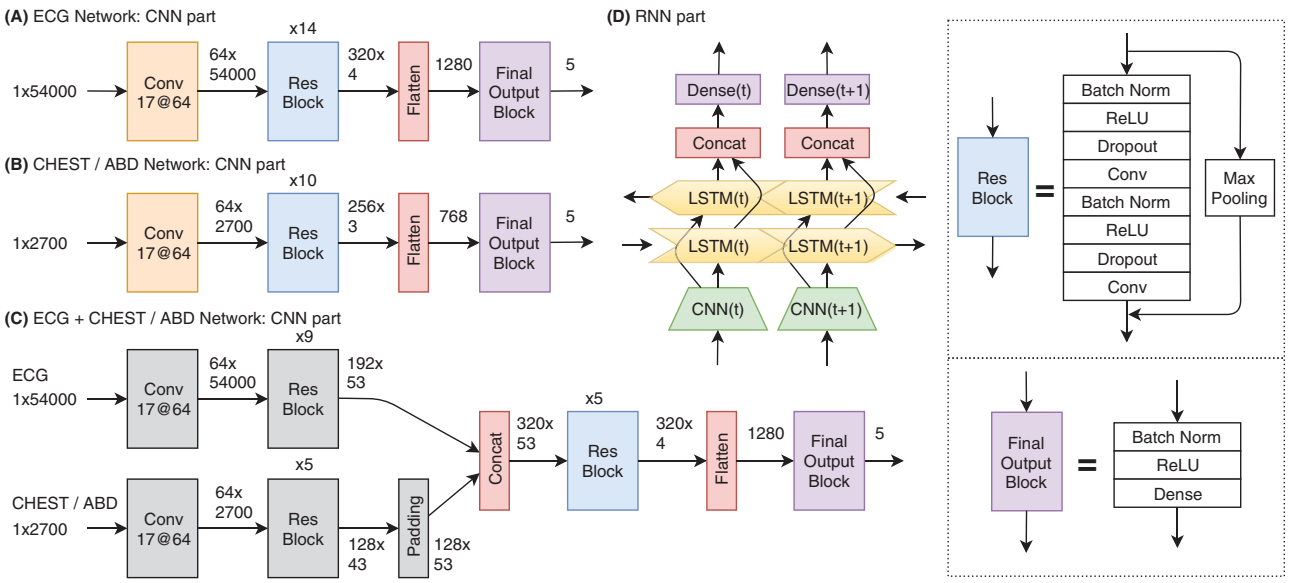

Figure 1. Deep neural network architecture. (A and B) CNN architecture using ECG, or CHEST or ABD as input. The numbers between blocks are the shapes of the output for each input 270-s time window. For example, "320 × 4" means 320 channels and four time points. "17@64" in the convolution layers means kernel size 17 points and 64 kernels. The repetition number of the residual blocks (Res Block) is marked above each block. Arrows indicate the flow of network activations. (C) The CNN architecture when using multiple signals as input. Gray blocks mean their weights are obtained from network trained in (A) and (B), then fixed during training the network. (D) RNN architecture, which uses the output from the CNN from every 270-s time window (corresponding to a 30-s time window). The output is fed into a bidirectional LSTM, followed by concatenation of the activations from both directions, and finally into a dense layer. The legends on the right show the detailed structure of the residual block and final output block. Inside each residual block, the first convolution layer subsamples the input by 4 (stride = 4) and the max pooling skip-layer connection also subsamples the input by 4.

stage imbalance), and confusion matrix as performance metrics. We show performance for staging five sleep stages according to AASM standards (W, N1, N2, N3, and R), and we additionally collapse these stages into three sleep super-stages, in two different ways. The first set of super-stages is "awake" (W) vs. "NREM sleep" (N1 + N2 + N3) vs. "REM sleep" (R); and the second set of super-stages is "awake or drowsy" (W + N1) vs. "sleep" (N2 + N3) vs. "REM sleep" (R).

To evaluate how many patients' data are needed to saturate the performance, we additionally trained the model multiple times with different numbers of patients and evaluated the performance. Specifically, for each fold, we randomly selected 10, 100, 1000, or all patients in the training folds, while keeping the testing fold unchanged. The reported performance metrics were based on the same held out testing set as used when training on all patients, ensuring results are comparable.

We obtained the 95% confidence intervals for Cohen's kappa using the formula in Cohen's original work [20], setting N as the number of unique patients; this represents the patient-wise confidence interval. For the macro-F1 score and weighted macro-F1 score, we obtained the 95% confidence interval by bootstrapping over patients (sampling with replacement by blocks of patients) 1000 times. The confidence interval was computed as the 2.5% (lower bound) and the 97.5% percentile (upper bound). Details about confidence interval computations are provided in the supplementary material.

Mathematically, the maximum possible value of Cohen's kappa for any number of categories is 1, indicating perfect agreement. However, practically, the maximum value of Cohen's kappa is the expert-to-expert inter-rater agreement (ee-IRA) where multiple experts perform sleep staging using ECG and/or respiration on the same data. ECG and/or respiration-based ee-IRA is not available in the literature because prior studies rely primarily on EEG to stage sleep. Therefore, we use the EEG-based ee-IRA as a practical upper bound for the maximum value of ee-IRA. Among 9 international institutes using 15 PSGs [21], for staging 5 stages, the EEG-based ee-IRA is 0.63 (95% CI 0.62–0.63) by Cohen's kappa; while it is 0.78 (95% CI 0.77–0.78) for staging 3 stages (Awake, NREM, REM). We computed relative Cohen's kappa, as the ratio of Cohen's kappa divided by the corresponding EEG-based ee-IRA, so that the performance becomes comparable between five stages and three stages.

Ranking sleep hypnogram features predictive of staging performance

Performance at the patient level, as measured by Cohen's kappa, varies across patients. To investigate potential sources of between-patient variation, we computed 25 features based on the sleep hypnogram, including the percentage of time in each of the five sleep stages (five features) and the percentage of time of each possible transition (W to N1, …, N3 to N2, 20 features). Since these features are correlated with each other, we used random forest regression, which is robust to collinearity, to learn relationship between these features and Cohen's kappa. Training was done by 10-fold cross-validation with no common patients between different folds to select the best hyperparameters, i.e. number of trees (from [20, 50]), max depth of each tree (from [5, 10]), and minimum number of patients at leaf nodes (from [5, 10]). We chose the best hyperparameters based on the average Pearson's correlation across all validation folds. We then refit the model using all the features and the best hyperparameters to arrive at the final model. We ranked the feature importance based on the average

decrease of impurity (variance) over all decision trees in the random forest by splitting on that feature. The sign of impact was based on the change of the predicted Cohen's kappa when perturbing each feature around its mean value, while fixing other features at their mean.

External validation

We used the Sleep Heart Health Study (SHHS) as an external validation dataset [22–25]. The dataset consists of two visits 5 years apart. The first visit contains 5793 patients, and the second visit is based on a subset of 2651 patients who returned after the first visit. Each recording has ECG, abdominal respiration, and chest respiration signals, as well as sleep stage annotations. We randomly selected 1000 recordings from both visits. Performance is reported using Cohen's kappa, macro-F1 score, and weighted macro-F1 score.

Results

Overall staging performance

In Figure 2, we show the confusion matrices for predicting all five sleep stages with different input signals, for both internal validation on MGH dataset and external validation on SHHS dataset. Using both ECG and ABD as input signals yields the best prediction results: for MGH, this network is correct in 78.9% of wake, 54.8% of N1, 68.4% of N2, 58.8% of N3, and 90.6% of REM time windows; for SHHS, this network is correct in 62.6% of wake, 59.1% of N1, 66.4% of N2, 43.4% of N3, and 88.2% of REM time windows. Most misclassifications are between W vs. N1, N1 vs. N2, and N2 vs. N3. For example, for MGH, 38.9% of N3 time windows are misclassified as N2, and 35.7% of N1 time windows are misclassified as either W or N2. This limitation is reduced when grouping the sleep stages as in Figure 3 and Supplementary Figure S2, so that both time windows of REM and NREM can be classified correctly with greater than 80% accuracy.

For each choice of input signal, we calculated Cohen's kappa, a statistic for assessing inter-reader agreement, and the standard deviation of the 95% confidence intervals. Results are show for the MGH dataset in Table 2. For the MGH dataset, ECG + ABD has the highest kappa, with values of 0.585 (±0.017) (all five stages), 0.760 (±0.019) (Wake vs. NREM vs. REM), and 0.735 (±0.017) (W + N1 vs. N2 + N3 vs. REM). The macro-F1 score and weighted macro-F1 score and their 95% confidence intervals are also shown in Table 2. Results for the external validation dataset (SHHS) are similar (Table 4). The learning curve when trained with different number of patients is shown in Supplementary Figure S3.

The relative Cohen's kappa (Cohen's kappa divided by the EEG-based expert-expert inter-rater agreement) for MGH dataset is shown in Table 3. Overall, ECG + ABD achieves 92.9% for five stages and 97.4% for three stages (Wake vs. NREM vs. REM). On this relative scale, the performance for five stages vs. three stages becomes comparable.

Staging performance on different groups of patients

In Table 5, we show Cohen's kappa for different population groups in the testing set using ECG + ABD as the input signals. Patients with older age have reduced performance. The Cohen's kappa for different population groups using other

A. MGH ECG

	W	N1	N2	N3	R
W	70.9	18.5	3.88	2.07	4.72
N1	23	46.4	15.7	3.44	11.4
N2	4.7	15.8	52.9	19.3	7.36
N3	2.25	2.81	24.1	69.2	1.69
R	4.71	4.22	4.73	1.71	84.6

B. MGH CHEST

	W	N1	N2	N3	R
W	73.3	21.2	3.5	0.231	1.81
N1	20.3	53.5	16	0.6	9.52
N2	2.91	17.1	66.4	6.43	7.15
N3	1.16	2.04	60	34.8	2.04
R	1.9	5.19	5.82	0.625	86.5

C. MGH ABD

	W	N1	N2	N3	R
W	75.2	19	3.53	0.299	1.98
N1	19.6	52.9	17	0.699	9.92
N2	2.68	15.5	65.6	8.48	7.73
N3	0.87	1.63	52.7	42.1	2.76
R	1.98	3.97	5.71	0.781	87.6

D. MGH ECG + CHEST

	W	N1	N2	N3	R
W	78.4	17.7	2.2	0.311	1.38
N1	20.2	55	14.9	0.754	9.1
N2	1.96	14.6	66.4	11.7	5.4
N3	0.677	1.3	34.1	63.1	0.822
R	1.16	3.53	3.59	0.36	91.4

E. MGH ECG + ABD

	W	N1	N2	N3	R
W	78.9	16.7	2.62	0.448	1.34
N1	19.5	54.8	16.2	0.879	8.62
N2	1.66	14.1	68.4	11	4.75
N3	0.518	1.07	38.9	58.8	0.74
R	0.975	3.57	4.46	0.406	90.6

F. SHHS ECG (external validation)

	W	N1	N2	N3	R
W	74.5	12.2	4.9	3.33	5.06
N1	23.9	45.3	16.5	2.97	11.4
N2	12	14.3	44.7	22.6	6.45
N3	7.73	3.34	21.6	66	1.39
R	12.2	5.94	6.41	3.41	72.1

G. SHHS CHEST (external validation)

	W	N1	N2	N3	R
W	45.2	23.8	20.8	3.79	6.47
N1	8.73	49.8	27.3	2.13	12.1
N2	3.87	14.3	63.4	8.98	9.45
N3	2.54	3.13	62	28.8	3.52
R	3.38	6.54	13	2.39	74.7

H. SHHS ABD (external validation)

	W	N1	N2	N3	R
W	48.7	20.4	14	5.18	11.6
N1	8.42	49.7	21.8	1.89	18.1
N2	2.49	15.9	63	6.26	12.3
N3	1.33	3.83	69	20.6	5.24
R	1.07	4.3	12.6	3.15	78.9

I. SHHS ECG + CHEST (external validation)

	W	N1	N2	N3	R
W	64.4	16.6	9.65	3.53	5.79
N1	13.2	60	17.6	0.779	8.49
N2	2.92	16.7	65.7	9.01	5.74
N3	1.54	3.69	54.6	39.2	1.03
R	1.99	6.24	9.31	0.767	81.7

J. SHHS ECG + ABD (external validation)

	W	N1	N2	N3	R
W	62.6	15.1	7	8.65	6.57
N1	11.7	59.1	16.7	0.779	11.7
N2	2.02	15.2	66.4	9.56	6.83
N3	0.891	2.58	52	43.4	1.11
R	0.728	4.04	5.76	1.28	88.2

Legend:

Stage by Technician

Network Prediction

Figure 2. Five-stage classification confusion matrices, comparing staging by sleep technicians vs. network predictions on the testing set for different input signals. Each row in the confusion matrix is the sleep stage annotated by the technician, while each column is the network prediction. The numbers are percentages. The top two rows are based on the pooled MGH testing set from all folds. The bottom two rows are based on the external validation SHHS dataset.

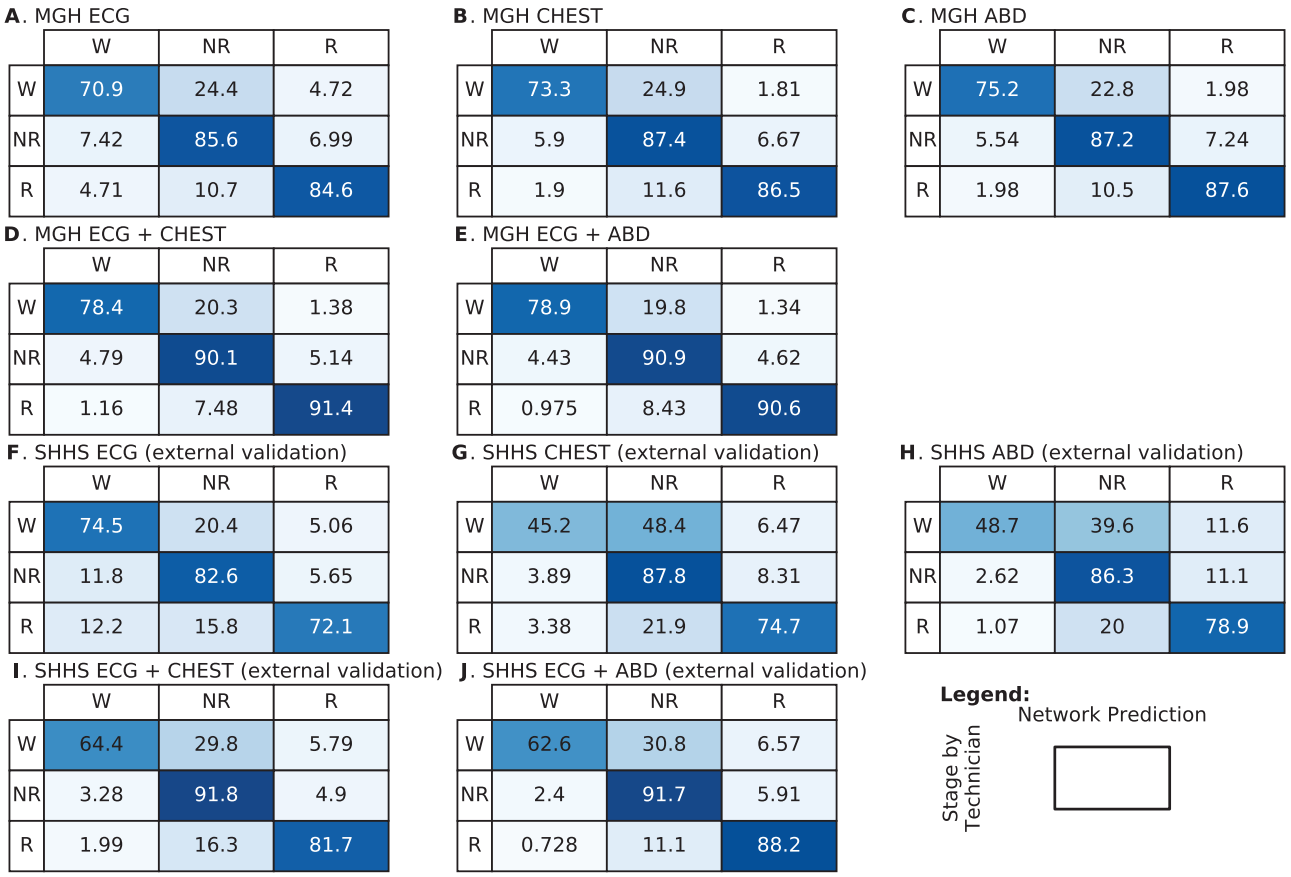

Figure 3. Three-stage classification confusion matrices, comparing staging by sleep technicians vs. network predictions on the testing set for different input signals. The three "super-stages" here are: "awake" (W) vs. "NREM sleep" (N1+N2+N3) vs. "REM sleep" (R). The top two rows are based on the pooled MGH testing set from all folds. The bottom two rows are based on the external validation SHHS dataset.

Table 2. Testing performances using different input signals and metrics.

			3 stages	
Metric	Input signal	5 stages	**W+N1 vs. N2+N3 vs. R**	**W vs. NR vs. R**
Cohen's kappa	ECG	0.490 (±0.017)	0.637 (±0.018)	0.646 (±0.021)
	CHEST	0.515 (±0.017)	0.681 (±0.017)	0.686 (±0.020)
	ABD	0.529 (±0.017)	0.687 (±0.017)	0.693 (±0.020)
	ECG + CHEST	0.586 (±0.017)	0.730 (±0.017)	0.750 (±0.019)
	ECG + ABD	0.585 (±0.017)	0.735 (±0.017)	0.760 (±0.019)
Macro F1 score	ECG	0.602 (0.598–0.606)	0.765 (0.762–0.769)	0.763 (0.759–0.767)
	CHEST	0.615 (0.611–0.618)	0.795 (0.792–0.798)	0.798 (0.895–0.801)
	ABD	0.629 (0.626–0.633)	0.797 (0.794–0.801)	0.802 (0.799–0.806)
	ECG + CHEST	0.681 (0.678–0.684)	0.829 (0.826–0.832)	0.836 (0.832–0.839)
	ECG + ABD	0.681 (0.678–0.684)	0.833 (0.831–0.836)	0.842 (0.839–0.845)
Weighted macro F1 score	ECG	0.615 (0.610–0.619)	0.792 (0.789–0.796)	0.842 (0.839–0.845)
	CHEST	0.642 (0.639–0.645)	0.812 (0.810–0.815)	0.854 (0.852–0.857)
	ABD	0.652 (0.649–0.655)	0.817 (0.814–0.820)	0.858 (0.856–0.860)
	ECG + CHEST	0.699 (0.696–0.702)	0.848 (0.846–0.851)	0.891 (0.889–0.893)
	ECG + ABD	0.701 (0.698–0.704)	0.852 (0.850–0.854)	0.897 (0.895–0.899)

input signals are shown in Supplementary Tables S1–S4 in the supplementary material. It is noticeable that for patients with very high BMI (>35), who usually have a larger waist circumference, performance is comparable to other groups when using ECG + ABD or ABD alone, especially in the W vs. NREM vs. REM combination.

Staging performance on individual PSGs

In Figure 4, we show the histogram of Cohen's kappa for each individual PSG using both ECG and ABD as input. The results indicate a fair amount of heterogeneity between PSGs, where the lowest extreme has kappa values around 0 and the highest extreme has kappa values around 0.9. Using the random forest

model described in the Methods Section, the top two sleep hypnogram features that (partially) tend to lead to better Cohen's kappa are percentage of time spent in REM or Awake; and the top five sleep hypnogram features leading to worse Cohen's kappa are percentage of time spent in N2 or N1 or N3, and the number of transitions between N2 and N3.

Dependence on temporal precision of R-peak timing in the ECG

Devices may vary in terms of the precision with which R-peaks can be measured in the ECG, therefore it is of interest to know how robust our model is to noise in R peak times. To investigate robustness to R-peak time noise, we simulated various levels of precision in R-peak determination by adding Gaussian noise to jitter the measurements of R-peak times. In Supplementary Figure S4 of the supplementary material, we see that adding zero-mean Gaussian jitter to the R peaks causes performance to drop progressively as the standard deviation of the jitter increases. Nevertheless, performance is relatively robust up to a noise standard deviation of approximately 20 ms.

Signal examples

To gain some insight into differences in breathing and heart rhythms that the deep neural network is using to distinguish sleep stages, we show some example whole night recordings in Figures 5–7. These examples are selected as "typical," meaning that they have the closest Cohen's kappa compared to the overall kappa across the testing set. The 60-s signal examples in Panel C are the signals where the deep neural network assigns the highest probability to the correct sleep stage within the recording. We can see a visible correspondence between the spectrogram and the sleep stages, as well as regions of mismatch between the spectrogram and EEG-based sleep stage. For example, in Figure 7, around 2 h and 4.5 h, the spectrogram of heart rate variability shows loss of very low frequency power, which is classified by the network as N3, but the EEG-based sleep stages contain both N2 and N3. More illustrations of the trained deep neural networks are shown in Supplementary Figures S5–S13 of the supplementary material.

Table 3. The Cohen's kappa relative to EEG-based expert-expert inter-rater agreement

Input Signal	5 stages Cohen's kappa relative to the EEG-based ee-IRA 0.63	3 stages Cohen's kappa (W vs. NR vs. R) relative to the EEG-based ee-IRA 0.78
ECG	77.8% (±2.7%)	82.8% (±2.7%)
CHEST	81.7% (±2.7%)	87.9% (±2.6%)
ABD	84.0% (±2.7%)	88.8% (±2.6%)
ECG + CHEST	93.0% (±2.7%)	96.2% (±2.4%)
ECG + ABD	92.9% (±2.7%)	97.4% (±2.4%)

Discussion

We hypothesized that it is possible to accurately stage sleep based on the ECG and respiratory signals using deep learning. Our results largely confirm this. Our key findings are: (1) ECG and respiratory signals contain substantial information about sleep stages; (2) N2 and N3 are relatively indistinguishable by respiration, collapsing sleep stages results in better staging performance; (3) staging is robust across a wide range of typical sleep disorders like OSA and PLMS, and commonly used medications; and (4) reduced staging accuracy is observed with older age.

Previous studies with similar goals to ours are summarized in Table 6. Some prior studies have also sought to stage sleep from ECG and respiration using deep neural networks. However, these studies suffered from small sample sizes, limiting generalizability. Only one prior study used more than 100 participants for training and evaluation. The large sample size of training and testing sets in the present study provides more robust results compared to prior literature, and increasing generalizability when applied to heterogeneous/external populations.

Sleep staging based on ECG and respiration has lower performance compared to using EEG. We previously performed EEG-based sleep staging with a deep neural network trained on data from the same set of patients used in the present work. The performance measured by Cohen's kappa is 0.805 within the MGH dataset when using all six EEG channels; is 0.764

Table 4. Performance on the external SHHS dataset.

Metric	Input Signal	5 Stages	3 Stages	
			W+N1 vs. N2+N3 vs. R	W vs. NR vs. R
Cohen's kappa	ECG	0.476 (±0.040)	0.611 (±0.043)	0.632 (±0.044)
	CHEST	0.388 (±0.042)	0.556 (±0.046)	0.515 (±0.050)
	ABD	0.397 (±0.042)	0.553 (±0.045)	0.542 (±0.048)
	ECG + CHEST	0.514 (±0.040)	0.663 (±0.042)	0.688 (±0.043)
	ECG + ABD	0.533 (±0.040)	0.675 (±0.041)	0.697 (±0.042)
Macro F1 score	ECG	0.545 (0.538–0.552)	0.745 (0.738–0.754)	0.757 (0.748–0.765)
	CHEST	0.480 (0.471–0.488)	0.713 (0.704–0.721)	0.687 (0.676–0.696)
	ABD	0.469 (0.461–0.477)	0.702 (0.694–0.710)	0.692 (0.684–0.701)
	ECG + CHEST	0.575 (0.566–0.582)	0.785 (0.777–0.793)	0.797 (0.788–0.803)
	ECG + ABD	0.586 (0.579–0.593)	0.793 (0.786–0.800)	0.802 (0.795–0.809)
Weighted macro F1 score	ECG	0.614 (0.605–0.622)	0.766 (0.758–0.774)	0.789 (0.781–0.798)
	CHEST	0.565 (0.556–0.574)	0.736 (0.728–0.745)	0.723 (0.714–0.731)
	ABD	0.568 (0.560–0.577)	0.732 (0.725–0.740)	0.737 (0.729–0.745)
	ECG + CHEST	0.664 (0.655–0.672)	0.800 (0.792–0.807)	0.825 (0.819–0.831)
	ECG + ABD	0.674 (0.668–0.682)	0.806 (0.800–0.813)	0.829 (0.822–0.835)

Table 5. Cohen's kappa in different subgroups using ECG + ABD as input.

			3 stages	
Category	Group	5 stages	W+N1 vs. N2+N3 vs. R	W vs. NR vs. R
Age	Young: 18 ≤ Age < 40	0.618	0.760	0.785
	Middle: 40 ≤ Age < 60	0.587	0.740	0.765
	Old: Age ≥ 60	0.552	0.703	0.735
Sex	Male	0.585	0.729	0.758
	Female	0.583	0.743	0.765
Type of Test	Diagnostic	0.589	0.738	0.760
	Split night	0.580	0.717	0.765
	All Night CPAP	0.580	0.746	0.759
BMI (kg/m^2)	Underweight: BMI < 18.5	0.541	0.732	0.756
	Normal: 18.5 ≤ BMI < 25	0.585	0.737	0.759
	Overweight: 25 ≤ BMI < 30	0.587	0.735	0.762
	Moderately obese: 30 ≤ BMI < 35	0.590	0.738	0.768
	More or equal to severely obese: BMI ≥ 35	0.580	0.733	0.757
AHI	Normal: AHI < 5	0.592	0.759	0.770
(per hour)	Mild: 5 ≤ AHI < 15	0.583	0.736	0.756
	Moderate: 15 ≤ AHI < 30	0.579	0.712	0.754
	Severe: AHI ≥ 30	0.559	0.677	0.753
Periodic limb movement	Normal: PLM < 5	0.592	0.742	0.774
(per hour)	Mild: 5 ≤ PLM < 15	0.588	0.734	0.762
	Moderate: 15 ≤ PLM < 30	0.584	0.736	0.753
	Severe: PLM ≥ 30	0.563	0.716	0.743
Medication	Antidepressant	0.581	0.736	0.760
	Benzodiazepine	0.584	0.738	0.759
	Diabetic	0.582	0.737	0.759
	Herbal	0.588	0.743	0.768
	Hypertension	0.584	0.733	0.761
	Neuroleptic	0.552	0.711	0.750
	Opiate	0.583	0.732	0.761
	Neuroactive	0.581	0.736	0.759
	Systemic	0.585	0.737	0.762
	RLS/PLMS	0.584	0.739	0.760
	Sleeping	0.584	0.736	0.758
	Stimulant	0.591	0.738	0.772
	Z-drug	0.581	0.733	0.761

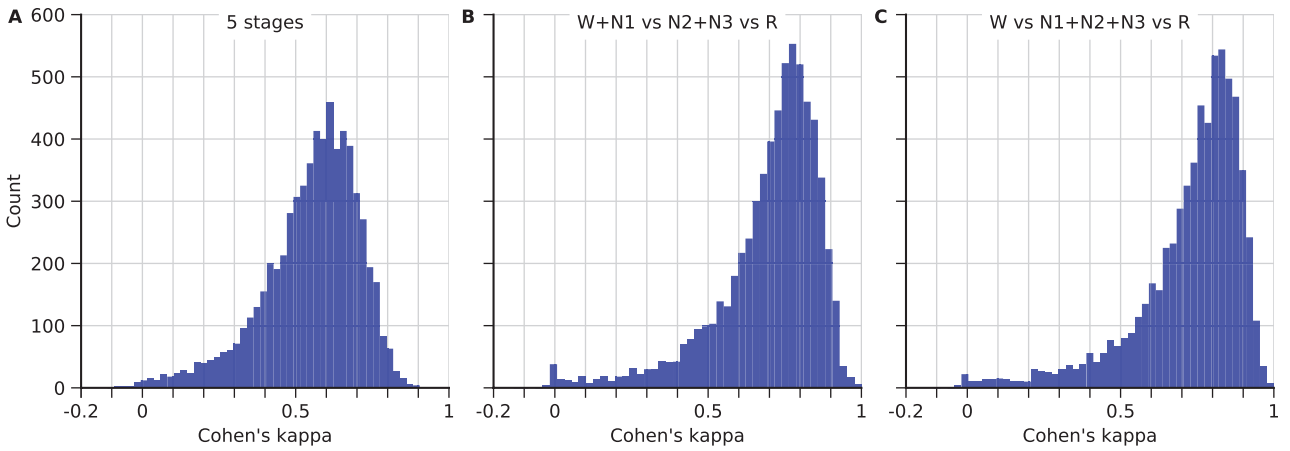

Figure 4. Histogram of Cohen's kappa values for individual PSGs using both ECG and ABD as input. The distributions are right-skewed.

within the MGH dataset when using only the two central EEG channels; and 0.732 when trained on MGH dataset and tested on SHHS dataset while restricting to the two central EEG channels. These two channel staging results are similar to human inter-rater agreement [16]. Note the human inter-rater agreement 0.63 used in [21] is measured based on scoring by nine international institutions; it is therefore not surprising that agreement is lower, though still higher than that using ECG and/or respiration. We only used the R-peaks as the input to our ECG network, therefore there is a reduction of information for ECG compared to respiration signals. This might partly explain why performance of ECG is least among all single-signal modalities.

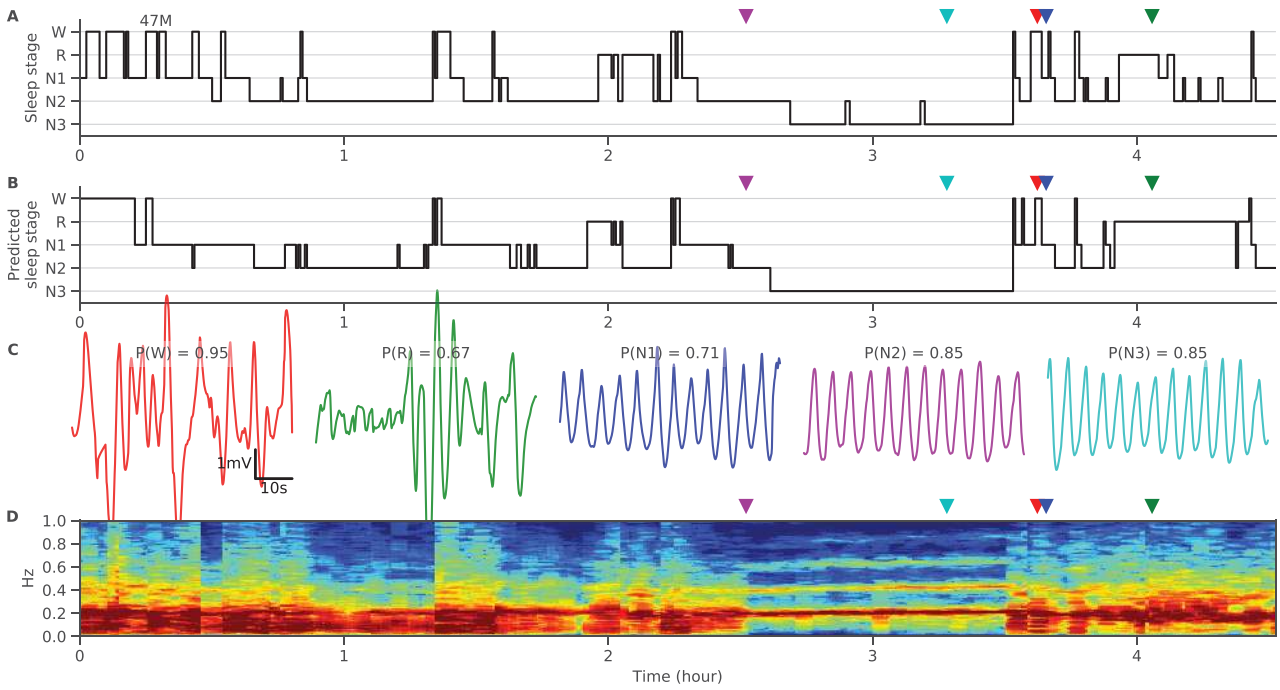

Figure 5. An example 47-year-old male. (A) The sleep stages over the whole night annotated by the technician (hypnogram). (B) The predicted sleep stages from the deep neural network using ABD respiration as input. (C) Example 60-s ABD segment from each sleep stage which is correctly classified and has the highest predicted probability of that stage. Different colors correspond to the triangle markers on other panels, which indicate the location of the example in the whole night recording. The number above each example signal indicates the probability of being that stage as predicted by the deep learning network. (D) The spectrogram of the ABD respiratory signal. The y-axis indicates the frequency.

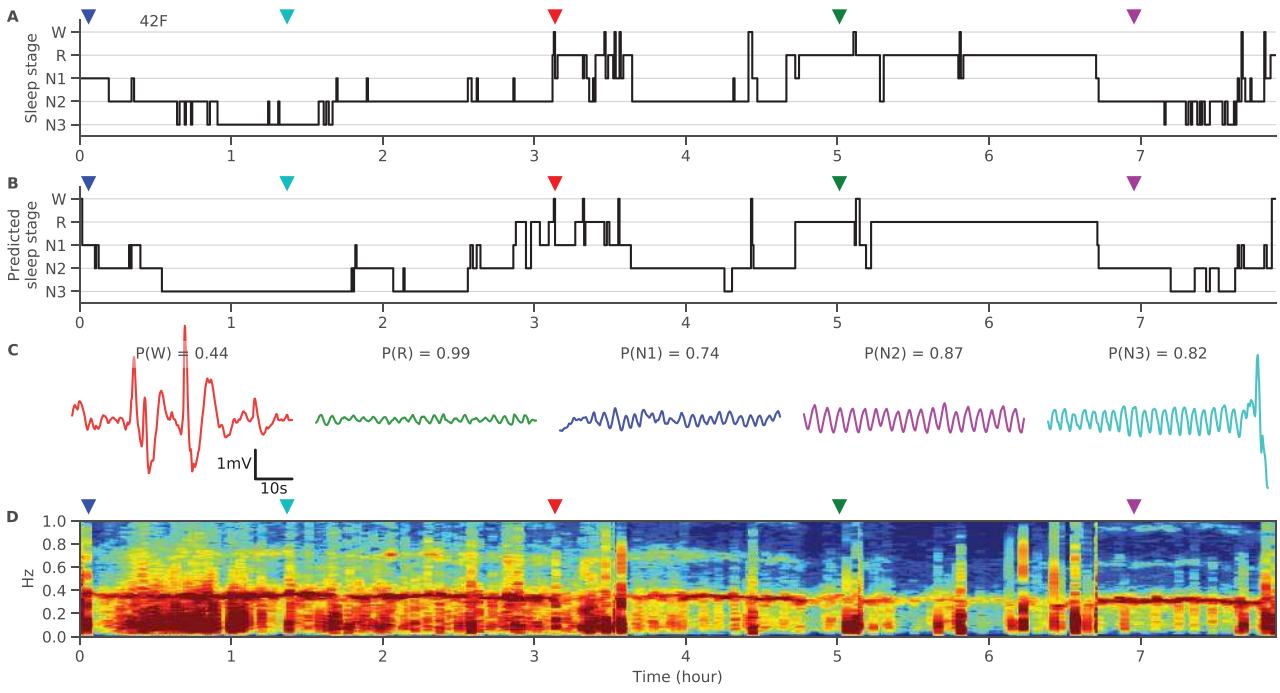

Figure 6. Similar to Figure 5, showing an example 42-year-old female using CHEST respiration as input. The scaling of the signals in (C) is the same as in Figure 5, but amplitude of these example signal itself happens to be smaller. It is possible that other time windows have larger or similar amplitude compared to Figure 5.

It is possible that using full-waveform ECG as the input might increase the ECG performance, at the cost of reducing robustness toward using ECGs from other (simpler) devices. We also did not use all three signals (ECG, ABD, and CHEST) here since our primary goal is to understand how well sleep staging can be done with a simple device. In our intended use, we would only have one or two, not all three signals.

The overall performance on the external validation SHHS dataset is reduced compared to that of inner validation (Figures 2 and 3, and Supplementary Figure S2). Notably, N1, N2, and REM

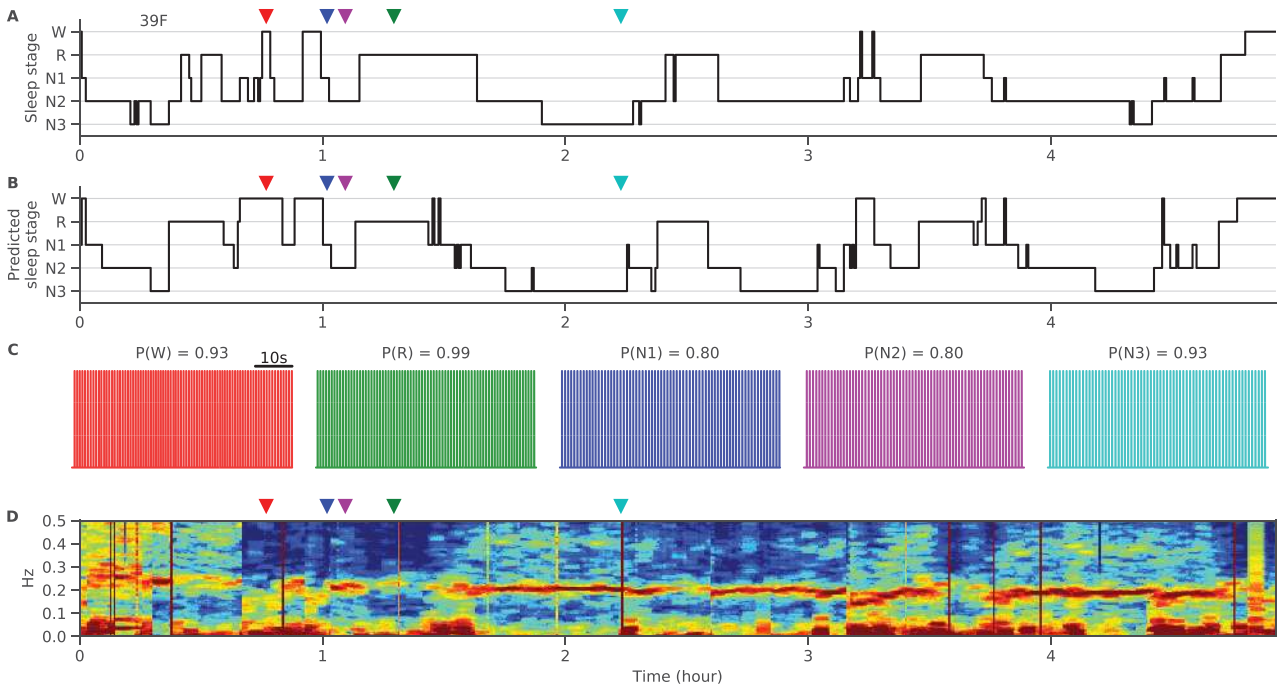

Figure 7. Similar to Figure 5, showing an example 39-year-old female using the R peaks from ECG as input. Panel C shows the R peaks represented as a binary sequence (see Methods). (D) The spectrogram of the R peak intervals.

Table 6. Related work in the literature.

Author (year)	Dataset size	Performance(κ is Cohen's kappa)	Type of Signal	Features
Sady et al. 2013 [11]	13 participants	3 stages (W, NREM, REM): accuracy = 78% 5 stages (W, N1, N2, N3, REM): accuracy = 62%	Photo-Plethysmogram, hemoglobin oxygen saturation, pneumotachograph	Heartbeat interval, time domain respiratory signals.
Long et al. 2014 [27]	48 participants	3 stages (W, NREM, REM): κ=0.48 4 stages (W, light sleep, deep sleep, REM): κ =0.41	Respiratory effort	Time domain, dissimilarity measure
Fonseca et al. 2015 [28]	48 participants	4 stages (W, light sleep, deep sleep, REM): κ = 0.49, accuracy = 69% 3 stages (W, NREM, REM): κ = 0.56 Accuracy = 80%	ECG + Respiratory inductance plethysmography	Time and frequency domain, nonlinear
Zhao et al. 2017 [29]	25 participants, 100 nights	4 stages (W, N1 + N2, N3, REM):κ = 0.70, accuracy = 79.8%	Radio frequency signal reflected off body (heartbeat, respiration)	Radio frequency spectrogram
Zhang et al. 2017 [30]	37,000 time windows	5 stages (W, N1, N2, N3, REM): Precision = 53.9% Recall = 56.0% Weighted macro-F1 score 53.2%	Heart rate derived from a wearable device	Frequency domain (DCT)
Radha et al. 2018 [31]	ECG: 352 participants, PPG: 60 participants	ECG: 6 stages (W, S1, S2, S3, S4, REM), κ = 0.61 and accuracy = 76.30% PPG: 5 stages (W, N1, N2, N3, REM), κ = 0.63 and accuracy = 74.65%	ECG and PPG	Selected features from time and frequency domain

stages have similar performance, while the reduction is mainly due to classifying awake and N3 stages. The reduction for awake stage could be due to the fact that SHHS data were collected at home, where the awake stage signal is noisier compared to sleep lab environment. Another interesting point is that the accuracy for awake stage using ECG only does not drop when validated on SHHS. This could be due to the fact that we were using R-peaks only, which are more robust in the case of noisy device (also see Supplementary Figure S4). N3 were mostly misclassified as N2 in SHHS to a greater extent compared to internal validation, emphasizing the intrinsic similarity between N2 and N3 in terms of respiration signals.

The improvement in staging performance when collapsing certain stages of sleep into super-stages may reflect information regarding the true biology of sleep states. N1 is an unstable transitional state with low probability and nondistinct EEG features.

About half of sleep is N2, and can show both stable and unstable characteristics, such as cyclic alternating pattern, apneic, or stable breathing in patients with sleep apnea. Different methods to characterize sleep depth and quality are available, and it will be important in future work to investigate whether further parsing of NREM sleep is meaningful using machine learning combined with methods such as the Odds Ratio Product of NREM sleep depth [26] or ECG-cardiopulmonary coupling [7]. An important future direction is to train the model from scratch for the combined sleep stages. Nevertheless, our current model has the flexibility to combine sleep stages as needed, while maintaining option of the granularity of five stage analysis.

The mild degradation of performance with age is not surprising when using conventional sleep stages as the ground truth. The reduction of N3 with age (mainly in males) is not accompanied by equal and simultaneous reductions in stable N2—thus, older individuals with equally reduced N3 may have very different N2 quality. In contrast, stable N2 and N3 may have very similar or identical cardiorespiratory signatures, making it difficult or impossible for deep learning models to reliably distinguish them. This is indeed validated by the result that more N2, N3, or the transitions between them is associated with worse staging performance. Thus, "errors" in discriminating these stages may reflect that EEG-based annotation in the reference standard is somewhat orthogonal to autonomic fluctuations.

The model is robust to different levels of AHI. We interpret this to mean that either autonomic features characteristic of stages are reasonably independent of sleep apnea, or more likely, that the network has learned normal, apneic, and other pathological patterns of the respiration signals change according to sleep stage. For example, REM and NREM interruptions in breathing may have distinct distributions of features such as event duration. The model is also robust of different levels of BMI, even using abdominal respiration in people with high BMI (BMI >35, likely to have large waist circumference).

Estimation of sleep states from cardiac and respiratory signals can simplify sleep tracking in health and disease. There are currently many wearable devices and in-bed devices with the capacity to record ECG and/or respiration. Our model should be able to be embedded in such devices to generate sleep parameters, such as NREM time, REM time, wake after sleep onset, facilitating high-quality sleep tracking. Another important scenario, which we are investigating now, is to monitor sleep in the ICU environment. Sleep quality is considered an important modifiable risk factor for the development of ICU delirium, yet at present sleep is either not monitored at all or is monitored only by asking the patient how well they slept. Nevertheless, most ICU patients have continuous ECG monitoring, and many have respiratory monitoring, thus our results lay a foundation for physiological monitoring of sleep in the ICU without imposing additional effort on nurses, and without introducing the need to additional monitoring devices.

Limitations of our analysis are as follows: (1) Our dataset includes only adults, and generalizability to the pediatric group will require additional study, (2) the 30-s based scoring of sleep limits the fine-grained analysis of sleep stages. This is especially true when sleep fragmenting conditions are present. Moreover, boundary zones may be amplified, such as transitions between wake-REM and NREM in the presence of sleep apnea in REM sleep. Such periods will introduce "error" in machine learning analyses, though these are biological features of sleep fragmentation rather than measurement or characterization error, such as arousal, apnea, or limb movement, and (3) due to the "black-box" nature of deep neural networks, there is limited insight into what the networks use as key features. For example, Supplementary Figures S11 and S12 suggest that the signal pattern within each breath may be important, rather than simply breath by breath variability. If true, this suggests that care may be needed when attempting to generalize these results to signals recorded by different devices, since some of the signal features could be altered depending on which devices is used. Future work to interpret what the networks have learned (beyond Figures 5–7, and Supplementary Figures S5–S13 in the supplementary material) is ongoing.

In conclusion, utilizing a large-scale dataset consisting of 8682 PSGs, we have developed a set of deep neural networks to classify sleep stages from ECG and/or respiration. ECG and respiratory effort provide substantial information about sleep stages. The best staging performance is obtained using both ECG and abdominal respiration. Staging performance depends on age, but is robust to body mass index and apnea severity.

Supplementary Material

Supplementary material is available at *SLEEP* online.

Data Availability

The trained networks and code to reproduce the figures are available in http://www.github.com/mghcdac/ecg_respiration_sleep_staging.

Acknowledgments

We gratefully acknowledge expert technical support from the Clinical Data Animation Center (CDAC) at Massachusetts General Hospital. There is a prior deposit of manuscript in a preprint data base (https://arxiv.org/abs/1908.11463).

Funding. M.B.W. received grants from National Institute of Neurological Disorders and Stroke (NIH-NINDS 1R01NS102190, 1R01NS102574, 1R01NS107291). R.J.T. receives grants from Category I American Academy of Sleep Medicine Foundation.

Conflict of interest statement. R.J.T. reports (1) patent and license to MyCardio, LLC, for ECG-spectrogram; (2) grant support, license, and intellectual property (licensed patents) from DeVilbiss-Drive Healthcare; (3) unlicensed patent for a device regulation CO_2 during positive pressure therapy, for central sleep apnea; (4) general sleep medicine consulting: GLG Councils, Guidepoint Global; (5) consultant-Jazz Pharmaceuticals. B.G. declares that the work was done while at Massachusetts General Hospital. He is currently a full time employee at Novartis Institutes of Biomedical Research with a role of Data Scientist.

References

1. Silber MH, *et al*. The visual scoring of sleep in adults. *J Clin Sleep Med*. 2007;**3**(2):121–131.
2. Chervin RD, *et al*. Respiratory cycle-related EEG changes: response to CPAP. *Sleep*. 2012;**35**(2):203–209.

3. Niizeki K, *et al*. Association between phase coupling of respiratory sinus arrhythmia and slow wave brain activity during sleep. *Front Physiol*. 2018;**9**:1338.
4. Penzel T, *et al*. Modulations of heart rate, ECG, and cardiorespiratory coupling observed in polysomnography. *Front Physiol*. 2016;**7**:460.
5. Thomas RJ, *et al*. Relationship between delta power and the electrocardiogram-derived cardiopulmonary spectrogram: possible implications for assessing the effectiveness of sleep. *Sleep Med*. 2014;**15**(1):125–131.
6. Iellamo F, *et al*. Baroreflex buffering of sympathetic activation during sleep: evidence from autonomic assessment of sleep macroarchitecture and microarchitecture. *Hypertension*. 2004;**43**(4):814–819.
7. Thomas RJ, *et al*. An electrocardiogram-based technique to assess cardiopulmonary coupling during sleep. *Sleep*. 2005;**28**(9):1151–1161.
8. Séi H. Blood pressure surges in REM sleep: a mini review. *Pathophysiology*. 2012;**19**(4):233–241.
9. Thomas RJ, *et al*. Cardiopulmonary coupling spectrogram as an ambulatory clinical biomarker of sleep stability and quality in health, sleep apnea, and insomnia. *Sleep*. 2017;**41**(2). doi:10.1093/sleep/zsx196
10. Bianchi MT. Sleep devices: wearables and nearables, informational and interventional, consumer and clinical. *Metabolism*. 2018;**84**:99–108.
11. Sady CC, *et al*. Automatic sleep staging from ventilator signals in non-invasive ventilation. *Comput Biol Med*. 2013;**43**(7):833–839.
12. Migliorini M, *et al*. Automatic sleep staging based on ballistocardiographic signals recorded through bed sensors. *Conf Proc IEEE Eng Med Biol Soc*. 2010;**2010**:3273–3276.
13. Tal A, *et al*. Validation of contact-free sleep monitoring device with comparison to polysomnography. *J Clin Sleep Med*. 2017;**13**(3):517–522.
14. Zaffaroni A, Doheny EP, Gahan L, *et al*. Non-Contact Estimation of Sleep Staging. In: Eskola H, Väisänen O, Viik J, Hyttinen J, eds. *EMBEC & NBC 2017*. Singapore Springer; 2017: 77–80.
15. Watson PL, *et al*. Atypical sleep in ventilated patients: empirical electroencephalography findings and the path toward revised ICU sleep scoring criteria. *Crit Care Med*. 2013;**41**(8):1958–1967.
16. Biswal S, *et al*. Expert-level sleep scoring with deep neural networks. *J Am Med Inform Assoc*. 2018;**25**(12):1643–1650.
17. Pan J, *et al*. A real-time QRS detection algorithm. *IEEE Trans Biomed Eng*. 1985;**32**(3):230–236.
18. Rebergen DJ, *et al*. ADARRI: a novel method to detect spurious R-peaks in the electrocardiogram for heart rate variability analysis in the intensive care unit. *J Clin Monit Comput*. 2018;**32**(1):53–61.
19. Hannun AY, *et al*. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;**25**(1):65–69.
20. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measure*. 1960;**20**(1):37–46.
21. Magalang UJ, *et al*.; SAGIC Investigators. Agreement in the scoring of respiratory events and sleep among international sleep centers. *Sleep*. 2013;**36**(4):591–596.
22. Dean DA 2nd, *et al*. Scaling up scientific discovery in sleep medicine: the national sleep research resource. *Sleep*. 2016;**39**(5):1151–1164.
23. Zhang GQ, *et al*. The national sleep research resource: towards a sleep data commons. *J Am Med Inform Assoc*. 2018;**25**(10):1351–1358.
24. Quan SF, *et al*. The sleep heart health study: design, rationale, and methods. *Sleep*. 1997;**20**(12):1077–1085.
25. Sleep Heart Health Research Group. Methods for obtaining and analyzing unattended polysomnography data for a multicenter study. *Sleep*. 1998;**21**(7):759–767.
26. Younes M, *et al*. Odds ratio product of sleep EEG as a continuous measure of sleep state. *Sleep*. 2015;**38**(4):641–654.
27. Long X, *et al*. Measuring dissimilarity between respiratory effort signals based on uniform scaling for sleep staging. *Physiol Meas*. 2014;**35**(12):2529–2542.
28. Fonseca P, *et al*. Sleep stage classification with ECG and respiratory effort. *Physiol Meas*. 2015;**36**(10):2027–2040.
29. Zhao M, *et al*. Learning sleep stages from radio signals: a conditional adversarial architecture. In: Proceedings of the 34th International Conference on Machine Learning; Aug 6–11, 2017.29.
30. Zhang X, *et al*. Sleep stage classification based on multilevel feature learning and recurrent neural networks via wearable device. *Comput Biol Med*. 2018;**103**:71–81.
31. Radha M, *et al*. LSTM knowledge transfer for HRV-based sleep staging. arXiv:180906221. 2018.