ORIGINAL ARTICLE

Large-Scale Automated Sleep Staging

Haoqi Sun, PhD[1,2]; Jian Jia, PhD[3]; Balaji Goparaju, Master of Science[4]; Guang-Bin Huang, PhD[5]; Olga Sourina, PhD[2]; Matt Travis Bianchi, MD, PhD[4]; M. Brandon Westover, MD, PhD[4,*]

[1]Energy Research Institute @ NTU, Interdisciplinary Graduate School, Nanyang Technological University, 639798, Singapore; [2]Fraunhofer IDM @ NTU, Nanyang Technological University, 639798, Singapore; [3]School of Mathematics, Northwest University, Xi'an, Shaanxi, 710127 China; [4]Massachusetts General Hospital, Neurology Department, Boston, MA; [5]School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore.

Study Objectives: Automated sleep staging has been previously limited by a combination of clinical and physiological heterogeneity. Both factors are in principle addressable with large data sets that enable robust calibration. However, the impact of sample size remains uncertain. The objectives are to investigate the extent to which machine learning methods can approximate the performance of human scorers when supplied with sufficient training cases and to investigate how staging performance depends on the number of training patients, contextual information, model complexity, and imbalance between sleep stage proportions.

Methods: A total of 102 features were extracted from six electroencephalography (EEG) channels in routine polysomnography. Two thousand nights were partitioned into equal (n = 1000) training and testing sets for validation. We used epoch-by-epoch Cohen's kappa statistics to measure the agreement between classifier output and human scorer according to American Academy of Sleep Medicine scoring criteria.

Results: Epoch-by-epoch Cohen's kappa improved with increasing training EEG recordings until saturation occurred (n = ~300). The kappa value was further improved by accounting for contextual (temporal) information, increasing model complexity, and adjusting the model training procedure to account for the imbalance of stage proportions. The final kappa on the testing set was 0.68. Testing on more EEG recordings leads to kappa estimates with lower variance.

Conclusion: Training with a large data set enables automated sleep staging that compares favorably with human scorers. Because testing was performed on a large and heterogeneous data set, the performance estimate has low variance and is likely to generalize broadly.

Keywords: sleep stages, EEG, machine learning, big data.

Statement of Significance

In the Big Data era, sleep research has the opportunity to benefit from large and heterogeneous data sets. However, how the performance of automated sleep staging scales with the data set size remains uncertain. The investigation of these questions provides several important observations that can generalize broadly, including contextual information, classifier complexity, and the precision of validation performance. The results suggest that the factors contributing to heterogeneity, such as stage transition frequency, should be further explored to continue improving staging performance.

INTRODUCTION

Overnight polysomnography (PSG) with electroencephalography (EEG) is the primary diagnostic test for evaluating patients with sleep problems.[1] The current standard for sleep staging is visual data review by certified sleep technicians following standardized rules of the American Academy of Sleep Medicine (AASM),[1] largely based on the original Rechtschaffen and Kales (R&K) criteria.[2] Automated sleep staging, if rigorously proven to provide information equivalent to human scoring, could substantially increase the utility and reach of sleep analysis in medical and research settings, by saving time and overcoming problems of scorer dependence.

The performance ceiling for an automated method judged against AASM or R&K is set by the inter-rater reliability (IRR) between human scorers. In the case of epoch-by-epoch comparison, IRR is usually measured by Cohen's kappa for two scorers or Fleiss' kappa for more than two scorers.[3] These kappa values reflect the degree of agreement between scorers above the chance level. IRR between human raters is imperfect. For AASM, IRR was 0.76 among eight European centers[3] and 0.63 among nine international centers.[4] For R&K, it was 0.68 among eight European centers.[3] IRR may vary according to the clinical and pathophysiological complexity of the data set.

Many studies have reported computer programs that attempt to replicate human scoring in the literature.[5–17] However, there is a general lack of external validation: few studies have convincingly demonstrated that algorithm performance can generalize to "all comers". Most prior studies have analyzed data from fewer than 100 individuals. Insufficient sample size leads to several important limitations. Staging algorithms trained on small data sets generalize poorly, that is, they exhibit variable performance on new cases because small data sets do not allow algorithms to account for the wide within- and between-patient variability inherent in real-world clinical settings. In addition, for machine learning methods which require training, it is difficult to split small data sets into training and independent testing sets, a crucial step in validation that mitigates the risk of over fitting (over-confidence in performance).

In the Big Data era, sleep research has the opportunity to benefit from large and heterogeneous data sets.[18] Potential benefits include (1) an automated staging system trained on a large and diverse set of patients has the potential to be robust, in the sense of performing reliably on EEG data from new patients; (2) large data sets provide the opportunity to take full advantage of machine learning methods that have achieved human-level performance on other pattern recognition tasks such as computer vision[19]; (3) large data sets allow validation, that is precise (low variance) and accurate (unbiased) estimates of performance on large testing sets that are independent of the training data.

Here, we present an automated EEG-based sleep staging system developed on a clinical data set of 2000 patients, split into 1000 patients for training and the other 1000 patients for testing (validation). We investigate how the following factors affect the performance of automated sleep staging:

number of training patients, contextual information between epochs, model complexity, and imbalance between sleep stage proportions. We also study the variance in testing performance as a function of the number of patients in the testing set.

METHODS

Data Set

The Partners Institutional Review Board approved retrospective analysis of our database without requiring additional consent for use of the clinically acquired data. PSG was performed according to AASM practice standards and scored by experienced technologists. Six EEG leads were subjected to analysis: bilateral frontal (F3 and F4), central (C3 and C4), and occipital (O1 and O2). Deidentified PSG exports (European Data Format of raw signals) were randomly selected without regard to indication and divided into a training set of 1000 patients and a testing set of 1000 patients without overlap.

Table 1 shows the clinical characteristics of the training and testing sets, which were similar to each other at the group level but heterogeneous across individual variables that might directly or indirectly impact sleep staging, such as age, body mass index, apnea-hypopnea index, and type of study. Approximately two-thirds of each set were diagnostic PSGs, with the remaining evenly divided between split night and continuous positive airway pressure titrations.

Sleep Stage Labeling by Human Scorers

EEG signals were scored in nonoverlapping 30-second epochs according to AASM standards as one of five stages: wake (W), rapid eye movement (REM), non-REM stage 1 (N1), non-REM stage 2 (N2), and non-REM stage 3 (N3).[1] In total, seven scorers annotated the data set but one scorer per EEG recording. Of the 2000 patients, we found 1617 patients with matched scorer information. In the training set, scorer S1 annotated 412 patients; S2, 304 patients; S3, 56 patients; S4, 17 patients; S5, 10 patients; S6, 5 patients; and S7, 2 patients. In the testing set, the proportions were similar: scorer S1 annotated 432 patients; S2, 300 patients; S3, 36 patients; S4, 19 patients; S5, 10 patients; S6, 10 patients; and S7, 4 patients.

Table 1—Clinical Characteristics of the Training and Testing Sets.

Variable	Training set[a]	Testing set[a]
Age (years)	52 (40–62)	51 (40–62)
% Male	52.3	55.4
BMI	31 (26–35)	30 (26–35)
ESS	7 (4–12)	8 (4–12)
TST (minutes)	374.5 (329.3–417.3)	379.3 (329–421)
N1 (%)	13.7 (8.2–21.7)	13.3 (7.9–21.2)
N2 (%)	52.3 (44.5–60.1)	52.5 (44.4–60.5)
N3 (%)	15.4 (7.6–22.4)	15.7 (8.1–22.3)
REM (%)	14.7 (8.8–20.6)	15.1 (9.9–20)
Efficiency (%)	83.6 (73.2–90.0)	83.5 (73–90.4)
AHI (/hour)[b]	4.0 (1.2–16.2)	4.1 (1.8–15.1)
PLMS (/hour)	5.5 (1.6–15.7)	5.9 (1.7–16.4)

[a]Values are median (interquartile range), except sex.
[b]Using 4% desaturation criteria.
AHI = apnea-hypopnea index; BMI = body mass index; ESS = Epworth sleepiness scale; PLMS = periodic leg movement during sleep; REM = rapid eye movement.

Preprocessing and Feature Extraction

Artifact Removal

We prespecified voltage threshold to identify EEG epochs for exclusion. Epochs with absolute amplitude larger than 500 μV were removed to avoid EEG contamination by movement artifact; epochs with no EEG (0 μV for more than 5 seconds) were also removed. Epochs with a maximum spectral power larger than 3000 times the median spectral power were found to be typically contaminated by 60 Hz power line noise and were also removed. The total amount of data removed by these procedures was about 6.7%.

Feature Extraction

Features were extracted from each 30-second epoch in both time and frequency domains. Features from the time domain included line length, a measure of the amplitude and frequency of oscillations in the EEG[20]; kurtosis, which measures the presence of extreme values, arising for example from K-complexes and eye movements[21]; and sample entropy, which measures signal irregularity.

Frequency domain or "spectral" features were derived from the spectrogram of each epoch. Because the duration of each epoch (30 seconds) is relatively long compared to transient events such as sleep spindles and K-complexes (about 1second), the spectrum for the whole epoch may not capture these events. Therefore, we further segmented each 30-second epoch into 29 subepochs of 2 seconds long with 1-second overlap. For each 2-second subepoch, we used the multitaper method[22] to estimate the power spectral density (PSD).

Other than the multitaper method, there are several PSD estimation alternatives, such as the fast Fourier transform (FFT) and periodogram. The FFT applies directly to the EEG signal without considering its underlying randomness, producing noisy estimates of the PSD. The periodogram is the Fourier transform of the autocovariance of the EEG signal, but this estimate does not converge to the true PSD even given infinitely long signal. In contrast, the multitaper method overcomes these disadvantages. It utilizes multiple mutually orthogonal windows (tapers), the Slepian sequences, to produce multiple periodograms of the windowed signal, and then takes the average as the estimated PSD. The multitaper method presents an optimal bias-variance trade-off.[22]

To further reduce noise in spectral features, we averaged spectrograms from contralateral channels, that is, F3-M2 and F4-M1, C3-M2 and C4-M1, O1-M2 and O2-M1. For each averaged spectrogram, we extracted the 95th percentile (robust version of maximum), minimum, mean, and standard deviation (SD) of the relative band power over all 29 subepochs from three bands: delta (0.5–4 Hz), theta (4–8 Hz), and alpha (8–12 Hz), as well as their ratios: delta/theta, delta/alpha, and theta/

alpha, where a small positive number was added to the denominator during implementation to avoid dividing by zero. The kurtosis of the spectrogram in the delta, theta, alpha, and sigma (12–20 Hz) bands were also extracted to measure transient bursts such as sleep spindles.

There were 102 features in total, summarized in Table 2. All features were log-transformed using $\text{sign}(x)\log(|x|+1)$ to render the distributions closer to Gaussian. Finally, each feature was standardized to zero mean and unit SD across all epochs in each patient, so that the features were on the same scale.

Sleep Stage Classifier

We used the extreme learning machine (ELM)[23] to create an automated algorithm that classifies each EEG epoch into one of the five sleep stages. ELM is a feed-forward neural network with one input layer, one hidden layer, and one output layer. The input layer has $d = 102$ nodes, one for each EEG feature. The hidden layer has L nodes, with a hidden weight matrix $A \in \mathbb{R}^{d\times L}$ and bias $b \in \mathbb{R}^{L}$ in the connection from the input nodes. They are generated randomly and then fixed. The output layer has $K = 5$ nodes, where the one with the largest activation indicates the sleep stage classification result. The output nodes are connected to the hidden nodes with weight matrix $\beta \in \mathbb{R}^{L\times K}$ Mathematical details are given in the appendix.

ELM classifiers have several advantages in the Big Data context (here, a large number of sleep-wake epochs, N). First, compared to kernel-based classifiers such as support vector machines (SVM), ELM explicitly decomposes the kernel function by random projection, which avoids storing the kernel matrix (with size N^2) during training. For example, in our case, there were $N = 848{,}815$ epochs from the 1000 training patients. Assuming each floating point number requires 8 bytes, kernel-based SVM would require about $8N^2 = 5.8\times10^{12}\,B = 5.2\,\text{TB}$ to store the kernel matrix during training, whereas ELM requires about 8Nd = 661 MB to store the feature matrix. The large memory consumption of storing the kernel matrix in SVM can be reduced by computing the matrix elements on the fly, however, at the cost of longer computation time, which does not scale to large data sets as in our case. Second, ELM can utilize "extreme logistic regression",[24] which gives a conditional probability of each stage for each epoch, and the maximum probability can be interpreted as confidence. This is in contrast to random forest where the probability is obtained less directly, as the proportion of classes in the trees or leaf nodes. Finally, ELM has a closed form solution (Appendix, equation (2)), thus iteration can be avoided and training time remains manageable. This is in contrast to machine learning classifiers that must be trained iteratively, such as neural networks. In the Supplementary Material, we provide experimental comparisons with other classifiers in terms of both staging performance and testing time.

RESULTS

Automated Staging Improves With Increasing Numbers of Training Patients

We first investigated how the number of training patients influences sleep staging performance according to the bias-variance theory in statistics.[25] Here, bias, also known as model

Table 2—Extracted Features.

Domain	Feature	Number	Formula
Time	Line length	6	$(1/(N-1))\sum_{i=1}^{N-1}\lvert x_{i+1}-x_i\rvert$
	Kurtosis	6	
	Sample entropy	6	
Frequency	95th percentile, min, mean, standard deviation of relative delta band power	3 × 4	Delta power/total power between 0.5 Hz and 20 Hz
	95th percentile, min, mean, standard deviation of relative theta band power	3 × 4	Theta power/total power between 0.5 Hz and 20 Hz
	95th percentile, min, mean, standard deviation of relative alpha band power	3 × 4	Alpha power/total power between 0.5 Hz and 20 Hz
	95th percentile, min, mean, standard deviation of delta-theta power ratio	3 × 4	Delta power/theta power
	95th percentile, min, mean, standard deviation of delta-alpha power ratio	3 × 4	Delta power/alpha power
	95th percentile, min, mean, standard deviation of theta-alpha power ratio	3 × 4	Theta power/alpha power
	kurtosis of delta band spectrogram	3	Computed over the whole spectrogram (time-frequency domain) of each band
	kurtosis of theta band spectrogram	3	
	Kurtosis of alpha band spectrogram	3	
	Kurtosis of sigma band spectrogram	3	
Total		102	For each 30-second epoch

underfitting, refers to error arising from overly low model complexity of the classifier relative to the data complexity. Variance, also known as overfitting, refers to error due to overly high model complexity relative to the data complexity. For classifiers with high variance (overfitting), adding more training samples improves generalization ability and therefore testing performance. For classifiers with high bias (underfitting), adding training samples does not substantially improve performance on the testing set. Rather, increasing model complexity or reducing data complexity is necessary to improve performance.

We conducted experiments with increasing numbers of training patients, ranging from 10 to 1000. The total testing set consisted of 1000 patients, randomly chosen and fixed from the entire data set. The training set and testing sets had no patients in common. For each number of training patients, training was repeated five times with different training patients (except when all training patients are used) and different randomly generated hidden weights and bias in ELM. For each repetition, the testing Cohen's kappa was computed using the confusion matrix derived from all epochs pooled over all 1000 testing patients.

To assess statistical significance of improvements in testing performance, we compared the testing performance on adjacent numbers of training patients using the Mann-Whitney *U* test. The comparison was carried out on the kappa values for each testing patient from five repetitions, thus having 5000 samples.

ELM has two parameters that influence model training: *L*, the number of nodes in the hidden layer; and *C*, a regularization parameter. Larger values of *L* lead to higher model complexity. Larger values of *C* lead to more bias and less variance. In this subsection, ELM was trained using fixed values of $L = 2000$ and $C = 0$.

Figure 1A shows the training Cohen's kappa with increasing numbers of training patients, and the Cohen's kappa on the testing set, where the mean and SD were computed from the five repetitions (Figure 1B will be described in the next subsection). Based on the comparison of adjacent cases, the curves can be divided into two phases: in phase I, the number of training patients was less than 300 and adding more training patients led to significant improvement in the testing performance (*p*-values: 10 vs. 50: 1.1×10^{-243}, 50 vs. 100: 3.4×10^{-8}, 100 vs. 200: 6.9×10^{-7}, 200 vs. 300: 0.04, not significant in other cases, Mann–Whitney *U* test). In phase II, the number of training patients was more than 300 and adding more training patients did not yield further significant improvement.

In phase I, the training kappa was higher than the testing kappa. This indicated low bias and high variance (overfitting). In other words, the complexity of the classifier was high compared to the training data complexity, so that the classifier fit its extra complexity to noise rather than signal. Increasing from 10 to 300 training patients improved the testing kappa from 0.485 to 0.608.

The confusion matrices when trained with 10 and 300 patients are shown in Figure 2A and B, respectively. Rows in the confusion matrix represent sleep stages assigned by the human scorer, while columns represent stages assigned by the classifier. Figure 2 shows the repetition with testing kappa closest to the mean value over five repetitions.

Comparing Figure 2B to Figure 2A, increasing the training patients from $n = 10$ to $n = 300$, staging accuracy was improved for all stages except stage N3 (Mann–Whitney *U* test on 1000 accuracy values from each testing patient; *p*-value: W 6.6×10^{-96}, N1 1.6×10^{-124}, N2 3.7×10^{-82}, N3 9.7×10^{-4}, R 1.4×10^{-30}). Therefore, in phase I, the generalization ability of the classifier was limited by the amount of training data.

Phase II is of particular interest in the context of Big Data. Confusion matrices for the classifier trained with 300 versus 1000 patients are shown in Figure 2B and C. The improvement was relatively small as training patient number increases, compared with phase I (Mann-Whitney *U* test on 1000 accuracy values from each testing patient; *p*-value: W n.s., N1 4.3×10^{-3}, N2 n.s., N3 n.s., R 0.011). The saturation indicates high bias and low variance (underfitting) regime. In other words, when adding more patients, heterogeneity becomes a dominating factor, indicating that model complexity may be insufficient to describe the complexity of the data. Other factors potentially contributing to saturation and methods that further improve performance in phase II are described in subsequent sections.

Contextual Information Improves Sleep Staging Performance

The classifier developed so far assigns stages independent of other epochs. This approach occasionally leads to excessive transitions in the hypnogram relative to human scorer, as seen in the example shown in the middle panel of Figure 3. In contrast, human scorers have access to and may take into account contextual information from neighboring epochs.[1,26] Thus, we next explore whether using contextual information can further improve sleep staging performance on the testing set.

Contextual information from neighboring epochs can be accounted for using hidden Markov models (HMMs). Here, the hidden states are the "true" stages provided by human scorers in the training data, while the observations are the "noisy" stages predicted by the classifier on the same data. The HMM decoding algorithm[27] finds the most probable sequence of hidden states using the stage transition matrix (describing the probability of transitioning from one state to any other) and the emission matrix for the observations (describing the probability distribution of the stages assigned by the ELM classifier as a function of the underlying true stage). The transition matrix was estimated using the sleep stages in the labeled training data. The emission matrix was generated by counting the pairs of sleep stages labeled by human scorers and the classifier on the same epochs.

The predicted hypnogram after smoothing is shown in the bottom panel of Figure 3. By taking into account contextual information, on average, the HMM decoding algorithm improved staging performance for both the training and testing kappa, as shown in Figure 1B. For the case of 1000 training patients, the overall performance on the 1000 testing patients was significantly improved from 0.615 to 0.666 (confirmed by Mann–Whitney *U* test on kappa values of each testing patient and five repetitions; *p*-value 1.3×10^{-92}).

On the other hand, for 74 of the 1000 testing patients (7.4%), HMM smoothing led to worse performance. In Figure 4, we show the patient with the largest decline in the Cohen's kappa. This patient has highly fragmented sleep compared to the patient in

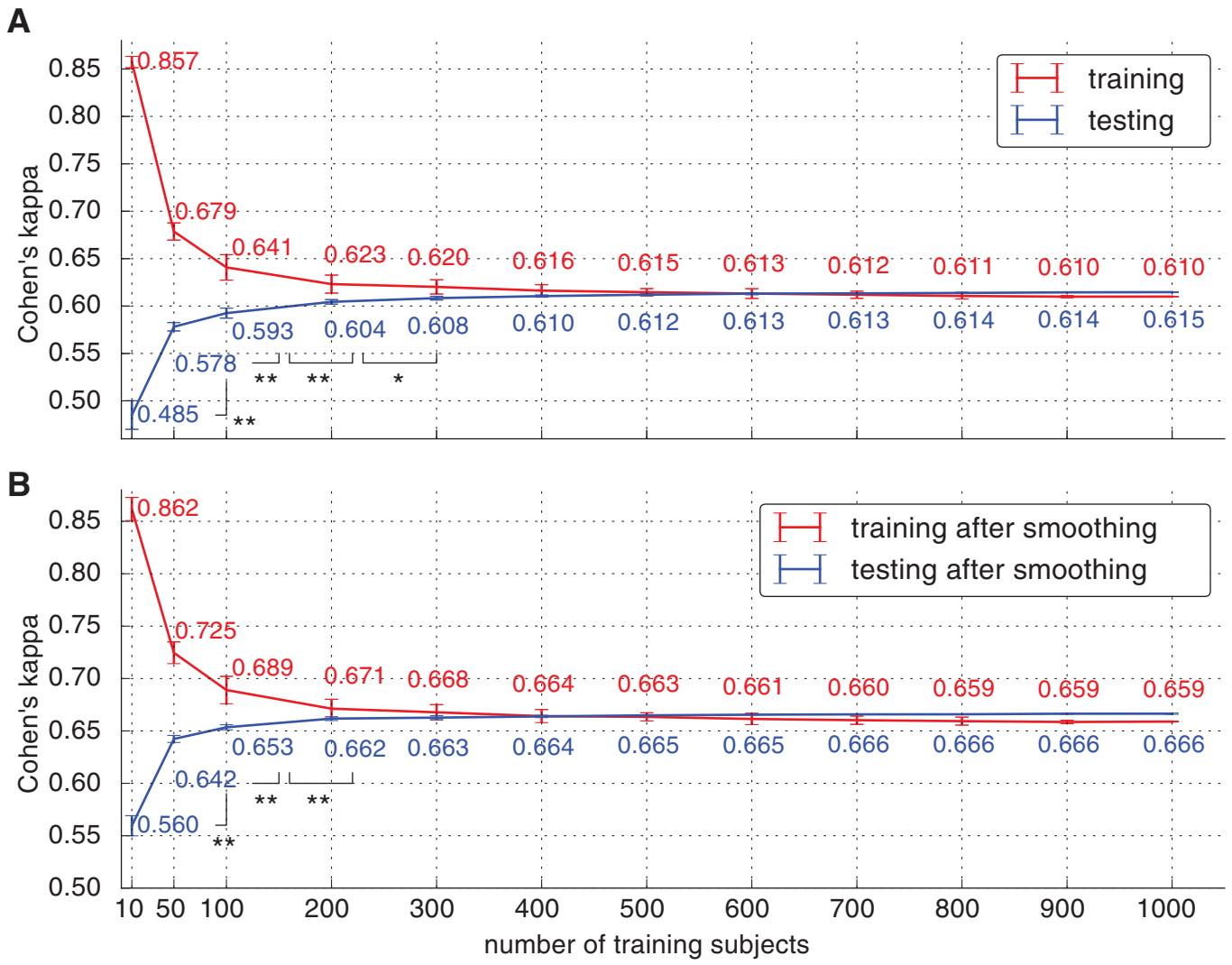

Figure 1—(A) Cohen's kappa for different numbers of training patients and the fixed 1000 testing patients. The testing Cohen's kappa was computed using the confusion matrix of all epochs from all 1000 testing patients. The mean and standard deviation of the kappa values from five repetitions are displayed. The ** or * markers indicate adjacent cases that exhibit a statistically significant improvement in the kappa statistic, based on Mann-Whitney *U* test (<0.01 or <0.05, respectively). (B) Cohen's kappa values after HMM smoothing, which will be described in the next subsection. HMM = hidden Markov model.

Figure 3. In this case, HMM smoothing tended to eliminate fragmentation and therefore did not capture this phenotype. A zoomed version of Figures 3 and 4 are provided in Supplementary Figure 2 in the supplementary material to further illustrate both cases.

Increasing Model Complexity Improves Sleep Staging Performance

Another way to further improve performance on testing patients in phase II, according to bias-variance theory, is to increase model complexity in order to meet data complexity. The model complexity of the ELM classifier is positively related to the number of hidden nodes L and negatively related to the regularization parameter C, as shown in Figure 5.

The confusion matrix obtained using $C = 0$ and $L = 20,000$ hidden nodes is shown in Figure 2D. The comparison to Figure 2C ($C = 0$ and $L = 2000$) confirms that more complex classifiers should be used to deal with increasing data complexity in the big data context (Mann–Whitney *U* test on 1000 accuracy values from each testing patient; *p*-value: W 0.016, N1 1.8×10^{-21}, N2 0.04, N3 n.s., R 8.7×10^{-6}).

Accounting for Stage Imbalance During Training Improves Performance

Stage N2 generally occupies 50% or more of total sleep time. During training, the fitting error for sleep stage N2 plays a dominating role, which makes the classifier best at predicting N2 but weaker at identifying other stages. Therefore, we next explored whether we could further improve testing performance by taking this imbalance into consideration while training the classifier.

Sleep stages can be balanced by weighting them differently in the objective function, as in the weighted ELM algorithm[28]; mathematical details are given in the Appendix. The resulting confusion matrix for the 1000 testing patients, when trained

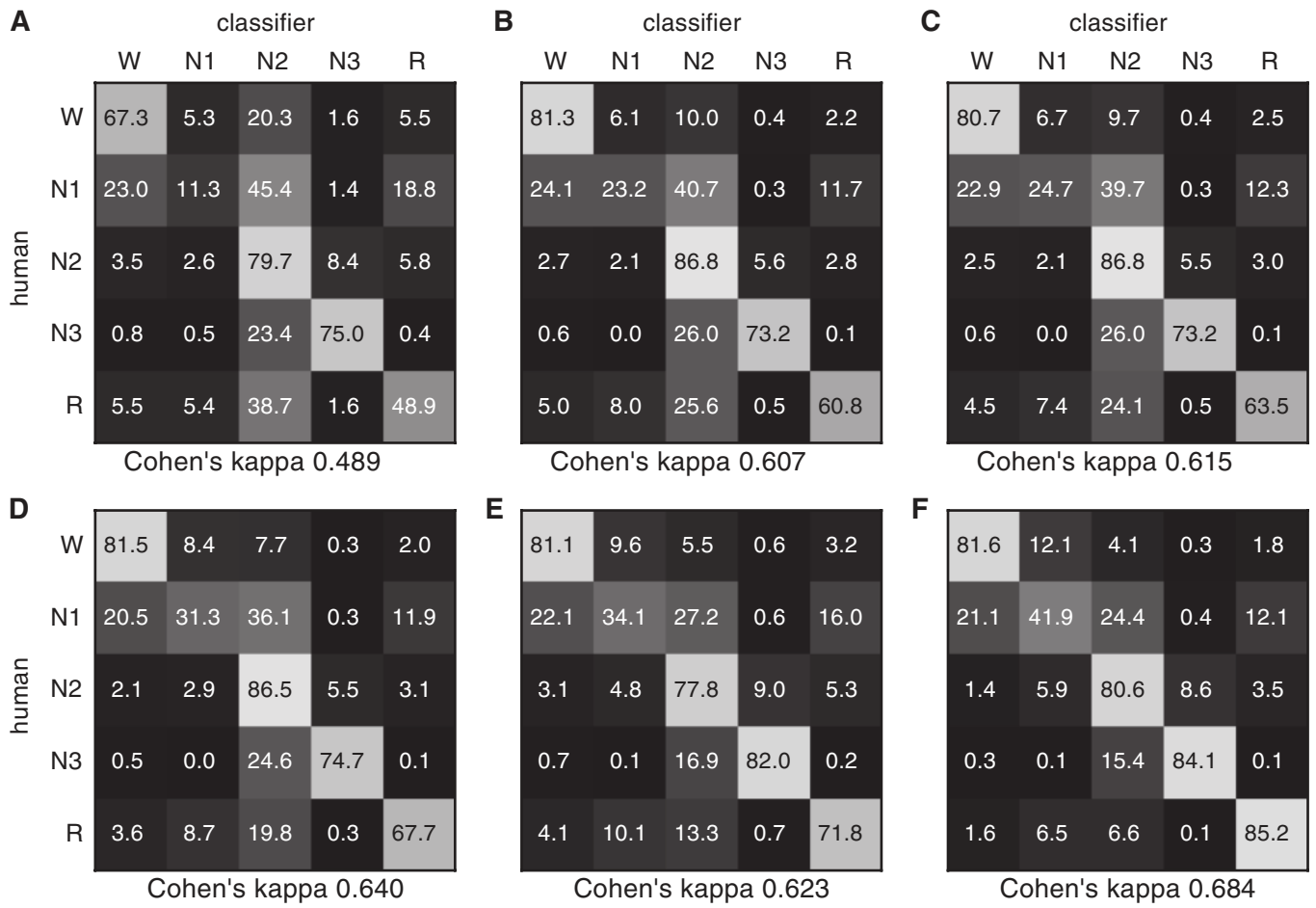

Figure 2—The confusion matrix of the fixed 1000 testing patients when trained with (A) 10 patients and 2000 hidden nodes; (B) 300 patients and 2000 hidden nodes; (C) 1000 patients and 2000 hidden nodes; (D) 1000 patients and 20,000 hidden nodes; (E) 1000 patients, 2000 hidden nodes and weighted training samples; and (F) 1000 patients, 20,000 hidden nodes, weighted training samples and smoothing. For each confusion matrix, the repetition with kappa value closest to the mean kappa value over five repetitions is shown. Values are given as percentages, which sum to 100 across rows (human scoring). The color is white for 100% and black for 0%.

with 1000 patients using weighted ELM is shown in Figure 2E. Comparing Figure 2E to Figure 2C (no weighting), the accuracy of sleep stages W, N1, N3, and R were improved; accuracy for stage N2 dropped since weighting does not change the model complexity of the classifier (Mann-Whitney U test on 1000 accuracy values from each testing patient; p-value: W n.s., N1 7.5×10^{-45}, N2 4.7×10^{-96}, N3 2.5×10^{-25}, R 1.3×10^{-17}).

Combining Strategies Maximizes Automated Staging Performance

The best performance was achieved by combining all 1000 training patients, a complex model (L = 20,000 hidden nodes, C = 0), compensation for class imbalance by weighted ELM training and contextualization via HMM smoothing. Based on epochs pooled from all 1000 testing patients in the repetition with testing kappa closest to the mean value over five repetitions, the overall Cohen's kappa was 0.684 and the accuracy was 76.9% (Figure 2F). In the following text, we use "overall" kappa to refer to this Cohen's kappa.

In Table 3, we compare human versus automated sleep parameters which commonly appear in clinical reports for the 1000 testing patients. While differences were detectable by statistical significance testing, which were unlikely to be clinically significant (with the large sample size, we were overpowered to detect small differences as significant). For all parameters, automated staging obtained values similar to those by human scorers.

The performance of automated staging varied across testing patients. The histogram of Cohen's kappa values for all 1000 testing patients is shown in Figure 6. Among the 1000 test patients, 523 had higher kappa values than the overall kappa 0.684. The tail at the left side was heavier than the right side (skewness = −1.2).

Evaluating Interscorer Difference in the Testing Set

In such a large data set, it is important to evaluate potential bias introduced by the sleep stages being scored by different human scorers. Specifically, we investigated whether the labels predicted by the classifier mimic the scoring of a particular scorer.

Since each patient recording was scored by only one scorer, we cannot evaluate the within-patient IRR. To give an estimate of IRR, we present the testing kappa from patients scored by

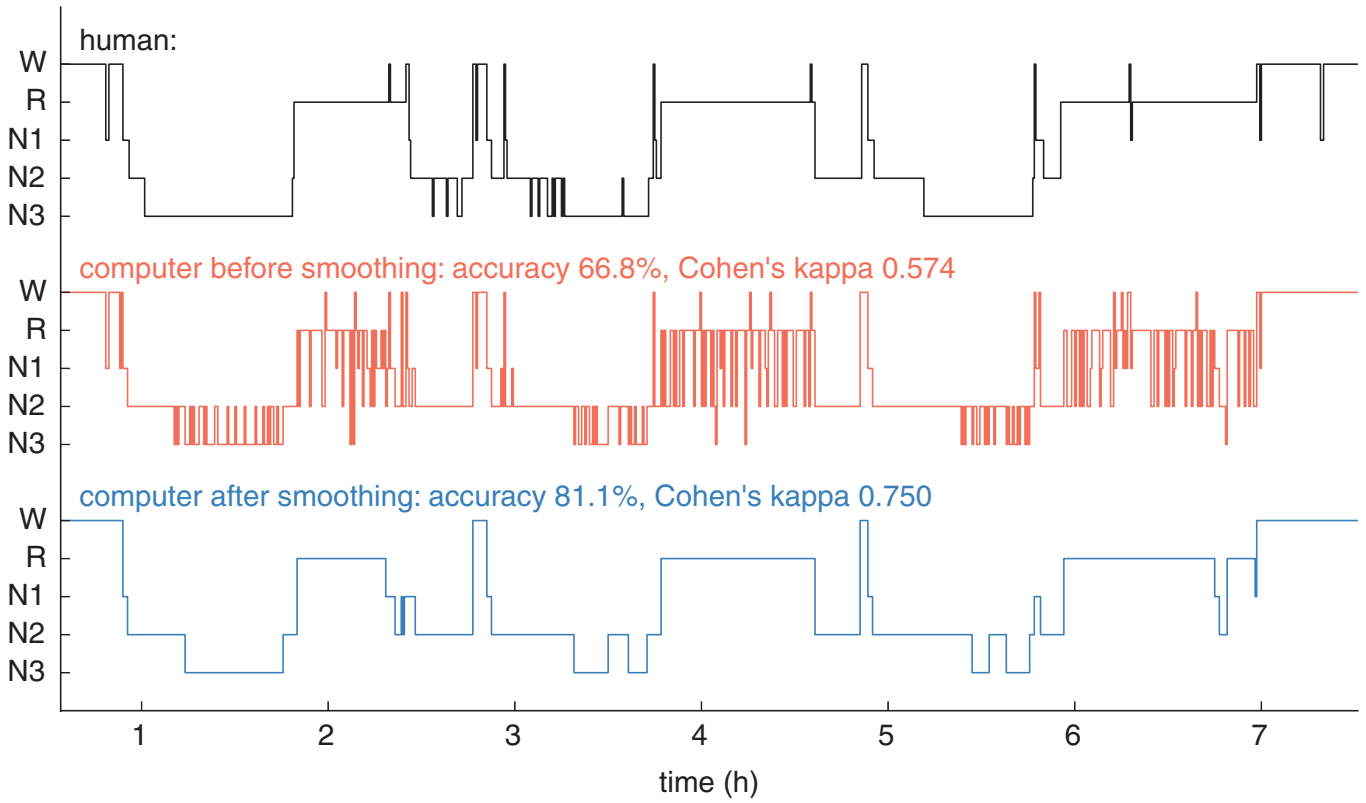

Figure 3—Hypnograms before and after HMM smoothing for the patient with largest post-smoothing improvement in the Cohen's kappa. HMM = hidden Markov model.

different scorers in Figure 7. We observed no specific relationship between the number of scored records per scorer (a proxy for experience) and the kappa value. In other words, the classifier does not appear to specialize in reproducing labels produced from scorers who scored the most cases. Scoring more patients does not mean the classifier learns different features or patterns. Applying Kruskal–Wallis *H* test, followed by post hoc Dunn's test with Bonferroni correction, the kappa value between classifier output and those labeled by S2 is significantly higher than that of S1 and S3. For the other scorers, the classifier matched their scores to a similar extent. On the other hand, if excluding scorer S2, the Kruskal–Wallis *H* test reveals no significant difference among the rest six scorers.

To further explain fluctuations in agreement among scorers, we checked the Spearman correlation between the median testing kappa (median line in each box in Figure 7) and the median value of various stage percentages in patients scored by the corresponding scorer. We found that only N1 is significant ($\alpha = 0.05$) with correlation −0.82. In other words, the percentage of scored N1 is negatively related to how well the classifier can learn from the scorer.

The Variance of Performance Estimates Reduces With the Size of the Testing Set

We have seen how the number of training patients affects sleep staging performance in Figure 1. Here, we investigate how the number of testing patients affects the variance of the performance estimate. To do this, the 1000 testing patients were first partitioned into nonoverlapping subsets, where four subsets were randomly selected. We computed four kappa values by pooling epochs from each subset. Then, we recorded the mean and SD of the four kappa values.

In Figure 8, we show the mean and SD of the testing kappa values, when each subset contains 5, 10, 15 … 250 patients. It is observed that the SD is negatively correlated with the size of the testing subsets (Spearman correlation −0.60, *p*-value 4.4×10^{-6}). In other words, smaller testing sets lead to estimates with higher variance (less precision). Thus, with a small testing set, we cannot be confident about the performance of a sleep staging algorithm. Precise validation requires a testing set of at least several hundreds of testing patients.

DISCUSSION

This study investigated the impact of four factors on automated sleep staging using a large-scale data set consisting of full-night EEG recordings from 1000 training and 1000 testing patients. Two strengths of our study are the large sample size and that the sample is a heterogeneous clinical population rather than a healthy or homogeneous population. Despite this heterogeneity, we find that a sufficiently complex machine learning algorithm, coupled with sufficiently large training data, is able to achieve testing Cohen's kappa comparable to those previously reported between human scorers (IRR).[3,4] Specifically, our best performing model achieved an IRR between human scorer and algorithm at 0.68, solidly in the range of prior reports of human scorer IRR values in the range 0.63–0.76.

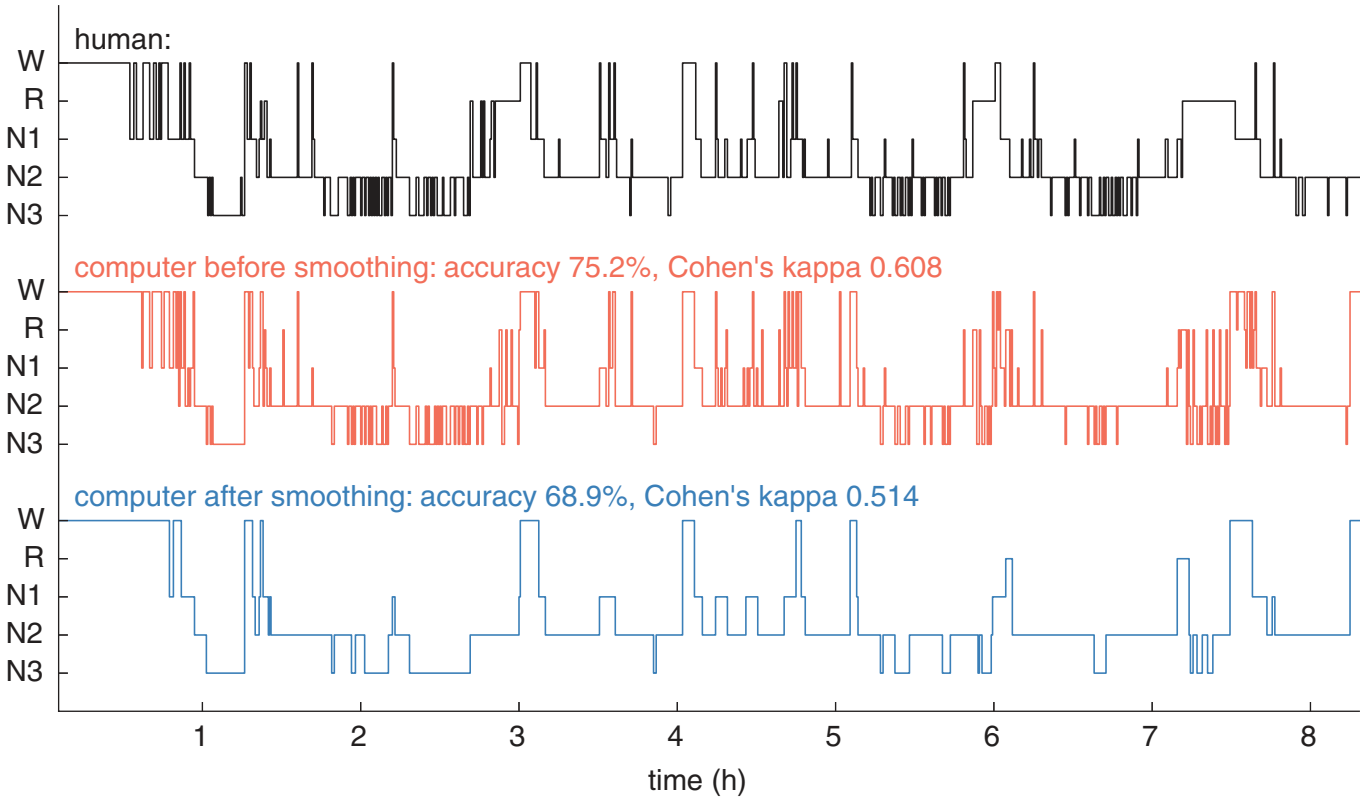

Figure 4—Hypnograms before and after HMM smoothing for the patient with largest post-smoothing decline in the Cohen's kappa. HMM = hidden Markov model.

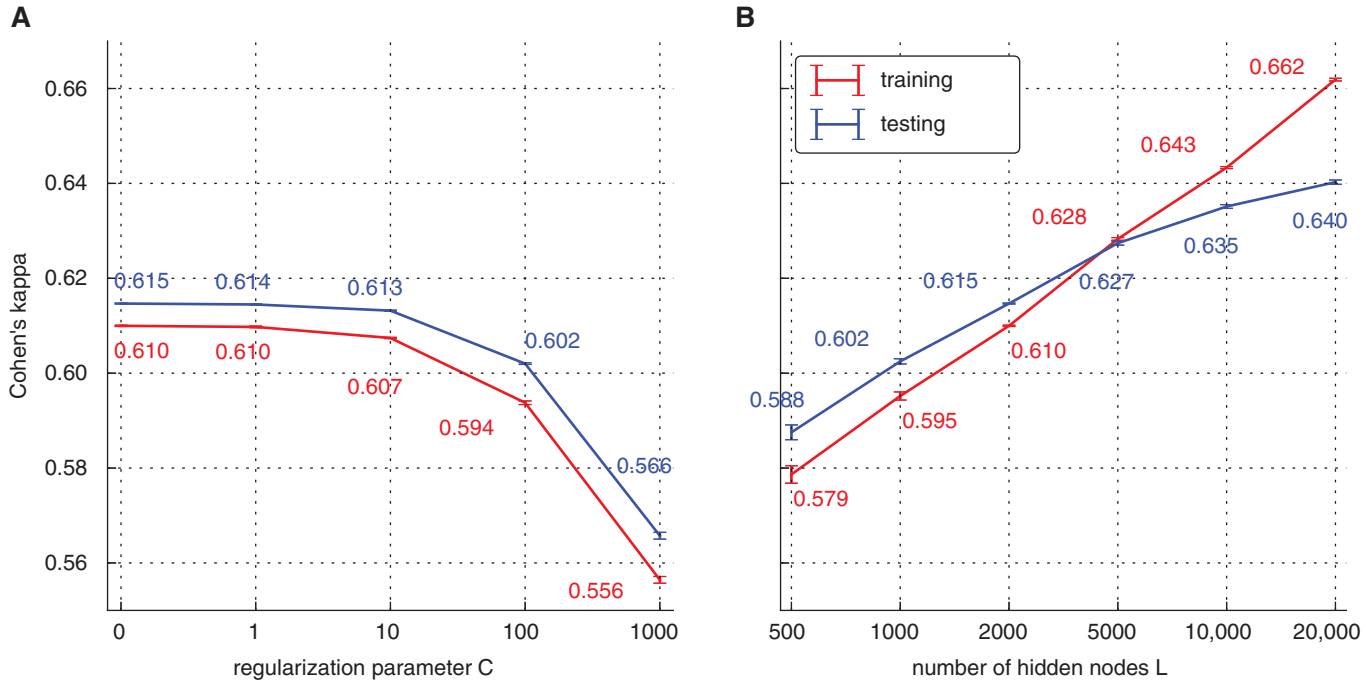

Figure 5—(A) Training and testing performance with different values of the regularization parameter C; (B) with different numbers of hidden nodes L. The mean and standard deviation of kappa values over five repetitions are shown in both panels.

We found that testing performance, as a function of training data size, can be divided into two phases (Figure 1). In phase I, the generalization ability of the classifier is limited by the amount of training data: sleep staging algorithms trained on small numbers of patients (here, <300) generalize poorly to test cases. Increasing the number of training cases improves staging performance from 0.485 to 0.608. In phase II, that is, with larger numbers of training patients, data complexity becomes a dominating factor: the complexity of the classifier limits its ability to fully model the complexity of the data, leading to diminishing returns with increasing size of the training set: increasing the number of patients from 300 to 1000 improves testing performance only from 0.61 to 0.62. To further improve testing performance, contextual information between epochs was used so that classification results were modified to account for neighboring epochs, which improved testing performance from 0.62 to 0.67.

We then increased model complexity by changing the parameters in the automated staging algorithm to allow greater model complexity (Figure 5) and accounted for class imbalance in the training data. The final testing performance considering all the above factors was 0.684 ± 0.0002, measured by Cohen's kappa based on the confusion matrix obtained from all epochs from all 1000 testing patients and 76.9% measured by accuracy.

When dealing with large-scale sleep data, limitations in automated scoring performance can be related to the noise and subjectivity of the scorers.[29] A trained classifier should satisfy two conditions. First, it should learn the general scoring pattern, or core features, across all scorers, instead of the pattern of a particular scorer. In Figure 7, the classifier indeed has similar agreement with all scorers, except a higher agreement with S2 compared to S1 and S3. To explain this, we did a correlation analysis which revealed that S2 tends to score less N1 than S1 and S3. This is consistent with the fact that the scorers have higher level of subjectivity when scoring N1, making N1 the stage with least agreement with the classification algorithm (Figure 2). Second, the trained classifier should have a kappa value independent of the number of patients scored by the scorer. This is evident in Figure 7, where the scorers are sorted in descending order of number of cases while the median line in each box does not follow descending order.

The large-scale data set also enables us to check the variance of the performance obtained on small testing sets (Figure 8). We found that performance estimates obtained from small-scale testing sets is subject to large variance. Similar observations have been reported previously in the study by Liang et al.[10] In their work, the performance on 17 testing subjects from the same data set reaches Cohen's kappa of 0.79 using decision tree. The same decision tree is then applied to another two subjects from a public sleep data set, has Cohen's kappa 0.68, exhibiting large variance.

Related Works

Some representative prior studies with less than 100 patients are summarized in Supplementary Table 1. Our results suggest that these studies may be subject to several important limitations due to small sample sizes: overfitting due to small training sets; high variance estimates of model performance due to small testing sets; and in several cases, limited scope due to testing only on healthy subjects rather than a heterogeneous population.

The largest prior study in the literature and hence the one most comparable to ours is the Somnolyzer 24 × 7 system,[6] which was validated using 590 recordings from the SIESTA project.[30] The SIESTA data set covers different genders, age groups, healthy

Table 3—Sleep Parameter Comparison Between Human Scorer and Classifier.

Sleep Parameter	Human		Classifier		Sig.[a]
	Mean	STD	Mean	STD	
Sleep latency (minutes)	23.4	23	25.3	23.6	
REM latency (minutes)	175.4	95.2	158.1	91.7	*
N1 latency (minutes)	24.4	24.8	27.7	27	*
N2 latency (minutes)	37.1	38.2	35.9	29.9	
N3 latency (minutes)	77.5	66.2	75.8	63.6	
Sleep efficiency (%)	83.5	14.1	83.1	11.3	
Total sleep time (minutes)	358.8	75.8	356.5	66.3	
NREM time (minutes)	302.7	62.7	294	51.6	*
REM time (minutes)	59.3	31.6	65.6	31.8	*
Awake time (minutes)	69.5	58.4	71.9	47.6	
N1 time (minutes)	54.9	37.3	46.4	25	*
N2 time (minutes)	189.5	60.9	181.7	45.2	*
N3 time (minutes)	62.5	36.6	69.1	28.3	*

*p-value < .05
[a]Significance after Holm-Bonferroni multiple test correction.
NREM = nonrapid eye movement; REM = rapid eye movement.

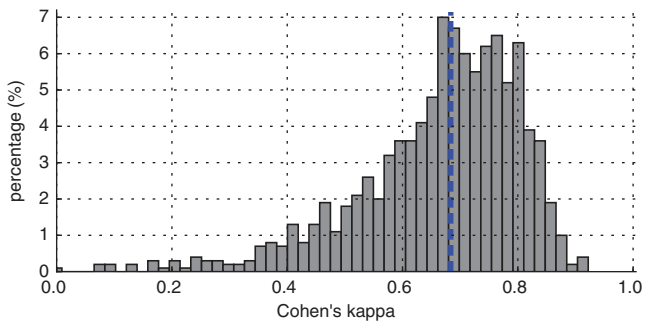

Figure 6—The histogram of the Cohen's kappa of each testing patient. The dashed line at 0.684 indicates the overall Cohen's kappa for all epochs pooled from all testing patients.

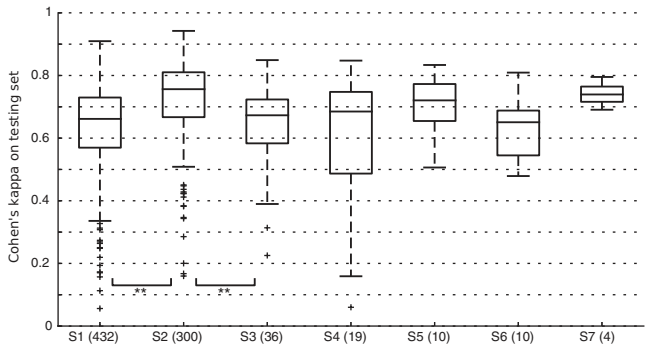

Figure 7—The testing Cohen's kappa for different scorers. The numbers in the x-axis labels are number of patients scored by the scorer. Kruskal-Wallis *H* test followed by post hoc Dunn's test suggests that the kappa values for S2 is higher than S1 and S3. Other scorers have similar kappa values. ** *p*-value < .01.

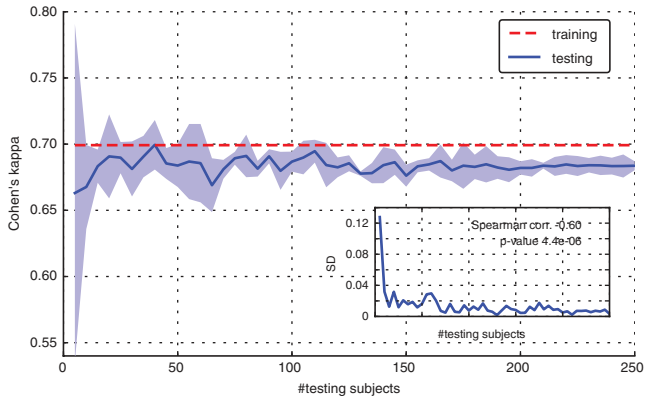

Figure 8—The precision of Cohen's kappa statistics for subsets of 1000 testing patients with different sizes. The red dashed line is the training Cohen's kappa. The blue solid line in the middle of the shading is the mean value of testing Cohen's kappa from four randomly selected nonoverlapping patient subsets. Shading indicates the area of ±standard deviation.

controls, and various sleep disorders and is labeled according to R&K scoring conventions. Various spectral features and spindle features were extracted from EEG, electrooculography, and electromyogram. An expert system was trained from half of the recordings and then tested on the other half of the recordings. Each node of the decision tree in the expert system had an linear discriminant analysis classifier with a predefined set of features. On the testing set, the Somnolyzer 24 × 7 obtained Cohen's kappa values of 0.69 and 0.67 when compared to two scorers, respectively, and 0.71 when compared to consensus scores. These results are comparable to our best model (Cohen's kappa = 0.68).

Another series of studies utilize the Michele Sleep Scoring System.[16,29,31–33] There are two such studies with more than 100 patients. Younes et al.[16] evaluated the performance of automated sleep scoring using frontal channels (F3 and F4) versus using central channels (C3 and C4) based on 102 PSG recordings. The recordings contain 27 healthy and heterogeneous clinical conditions such as 49 obstructive sleep apnea (OSA), 14 insomnia, and 23 periodic limb movements (11 overlaps with OSA). In Yonnes et al.,[31] the authors proposed odds ratio product as a continuous measure of sleep depth to augment the discrete sleep stages defined in R&K and AASM. The study was based on 58 training patients and 56 testing patients (total 114). Although different clinical conditions were included, both of the studies had fewer than 100 patients in each clinical condition, and thus, given the heterogeneity seen across sleep disorders categories, the external validity remains uncertain.

Limitations

Although we have validated our approach on a large and clinically heterogeneous data set, we have not performed validation outside a single center. It is thus possible that the population of patients seen in our center is not fully representative of populations elsewhere. Also, many factors contributing to heterogeneity could be further explored with even larger data sets, such as medication effects, medical comorbidities, specific sleep disorders of varying severity, age-specific effects, etc. It is possible that certain physiological EEG features would be more or less relevant across subpopulations.

APPENDIX

MATHEMATICAL DETAILS OF EXTREME LEARNING MACHINE

Suppose the features are stored in a matrix $X \in \mathbb{R}^{N\times d}$, where it has N rows (epochs) and each row contains $d = 102$ features from each epoch. The hidden layer activation $H \in \mathbb{R}^{N\times L}$ is computed using sigmoid function:

$$H = \text{sigmoid}(XA + b) = \frac{1}{1+\exp(-XA-b)}. \quad (1)$$

where $A \in \mathbb{R}^{d\times L}$ and $b \in \mathbb{R}^{L}$ are randomly generated and then fixed.

During training, the output weight vector β is computed as

$$\beta = \underset{\beta}{\text{argmin}} \frac{1}{2N}\|H\beta - Y\|_F^2 + \frac{C}{2L}\|\beta\|_F^2$$
$$= \left(H^{\mathrm{T}}H + \frac{CN}{L}I\right)^{-1} H^{\mathrm{T}}Y, \quad (2)$$

where $I \in \mathbb{R}^{L\times L}$ is an identity matrix; $X_F = \sqrt{\sum_{ij} X_{ij}^2}$ is the Frobenius matrix norm; $C \geq 0$ is a regularization parameter; and $Y \in \mathbb{R}^{N\times K}$ contains the training labels, i.e. the sleep stages labeled by the human scorer. Y is encoded in the "one-vs-all" style according to the extreme logistic regression algorithm,[24] where in each row the k-th element is 2 and others are −2 for an epoch belonging to sleep stage k, so that the output probability can be approximated by $\text{sigmoid}(H\beta)$.

ELM has two parameters: number of hidden nodes L and regularization parameter *C*. Larger *L* leads to higher model complexity since it is the dimensionality of the features in the hidden layer. Larger *C* leads to more bias and less variance, and vice versa.

MATHEMATICAL DETAILS OF WEIGHTED EXTREME LEARNING MACHINE

Sleep stages can be balanced by weighting them differently in the objective function. In weighted ELM,[28] each training epoch is associated with a weight, which forms a diagonal matrix $W \in \mathbb{R}^{N\times N}$. Formally, the output weight β is computed as

$$\begin{aligned}\beta &= \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \| \tilde{W}\left(H\beta - Y\right) \|_F^2 + \frac{C}{2L} \| \beta \|_F^2 \\ &= \left(H^T \tilde{W}^T \tilde{W} H + \frac{C}{L} I \right)^{-1} \tilde{W}^T \tilde{W}^T \tilde{W} Y \\ &= \left(H^T W H + \frac{C}{L} I \right)^{-1} H^T W Y, \end{aligned} \quad (3)$$

where we have used $W_{ii} = \left(1/\sqrt{N_k}\right) / \sum_i W_{ii}$, that is, the weight of epoch i was inversely proportional to the square root of the number of epochs belonging to sleep stage N_k, then normalized so that all weights sum to 1.

REFERENCES

1. Iber C. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology And Technical Specifications. American Academy of Sleep Medicine; 2007 IL.
2. Rechtschaffen A, Kales A. A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects. National Institutes of Health Publication no. 204; 1968.
3. Danker-Hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res. 2009; 18(1): 74–84.
4. Magalang UJ, Chen NH, Cistulli PA, et al.; SAGIC Investigators. Agreement in the scoring of respiratory events and sleep among international sleep centers. Sleep. 2013; 36(4): 591–596.
5. Schaltenbrand N, Lengelle R, Toussaint M, et al. Sleep stage scoring using the neural network model: comparison between visual and automatic analysis in normal subjects and patients. Sleep. 1996; 19(1): 26–35.
6. Anderer P, Gruber G, Parapatics S, et al. An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24 × 7 utilizing the Siesta database. Neuropsychobiology. 2005; 51(3): 115–133.
7. Berthomier C, Drouot X, Herman-Stoïca M, et al. Automatic analysis of single-channel sleep EEG: validation in healthy individuals. Sleep. 2007; 30(11): 1587–1595.
8. Anderer P, Moreau A, Woertz M, et al. Computer-assisted sleep classification according to the standard of the American Academy of Sleep Medicine: validation study of the AASM version of the Somnolyzer 24 × 7. Neuropsychobiology. 2010; 62(4): 250–264.
9. Fraiwan L, Lweesy K, Khasawneh N, Fraiwan M, Wenz H, Dickhaus H. Classification of sleep stages using multi-wavelet time frequency entropy and LDA. Methods Inf Med. 2010; 49(3): 230–237.
10. Liang SF, Kuo CE, Hu YH, Cheng YS. A rule-based automatic sleep staging method. J Neurosci Methods. 2012; 205(1): 169–176.
11. Malhotra A, Younes M, Kuna ST, et al. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. Sleep. 2013; 36(4): 573–582.
12. Lajnef T, Chaibi S, Ruby P, et al. Learning machines and sleeping brains: automatic sleep stage classification using decision-tree multi-class support vector machines. J Neurosci Methods. 2015; 250: 94–105.
13. Wang Y, Loparo KA, Kelly MR, Kaplan RF. Evaluation of an automated single-channel sleep staging algorithm. Nat Sci Sleep. 2015; 7: 101.
14. Punjabi NM, Shifa N, Dorffner G, Patil S, Pien G, Aurora RN. Computer-assisted automated scoring of polysomnograms using the somnolyzer system. Sleep. 2015; 38(10): 1555–1566.
15. Hassan AR, Bhuiyan MIH. A decision support system for automatic sleep staging from EEG signals using tunable q-factor wavelet transform and spectral features. J Neurosci Methods. 2016; 271: 107–118.
16. Younes M, Younes M, Giannouli E. Accuracy of automatic polysomnography scoring using frontal electrodes. J Clin Sleep Med. 2016; 12(5): 735–746.
17. Younes M, Soiferman M, Thompson W, Giannouli E. Performance of a new portable wireless sleep monitor. J Clin Sleep Med. 2017; 13(2): 245–258.
18. Dean DA II, Goldberger AL, Mueller R, et al. Scaling up scientific discovery in sleep medicine: the national sleep research resource. Sleep. 2016; 39(5): 1151–1164.
19. Taigman Y, Yang M, Ranzato MA, Wolf L. Deepface: closing the gap to human-level performance in face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014; Columbus, OH.
20. Esteller R, Echauz J, Tcheng T, Litt B, Pless B. Line length: an efficient feature for seizure onset detection. In Proceedings from the Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE; 2001; Istanbul, Turkey.
21. Zoubek L, Charbonnier S, Lesecq S, Buguet A, Chapotot F. Feature selection for sleep/wake stages classification using data driven methods. Biomed Signal Processing and Control. 2007; 2(3): 171–179.
22. Babadi B, Brown EN. A review of multitaper spectral analysis. IEEE Trans Biomed Eng. 2014; 61(5): 1555–1564.
23. Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: theory and applications. Neurocomputing. 2006; 70(1): 489–501.
24. Ngufor C, Wojtusiak J. Extreme logistic regression. Adv Data Anal Classif. 2016; 10(1): 27–52.
25. Bishop CM. Pattern Recognition and Machine Learning. New York: Springer; 2006.
26. Silber MH, Ancoli-Israel S, Bonnet MH, et al. The visual scoring of sleep in adults. J Clin Sleep Med. 2007; 3(2): 121–131.
27. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, UK: Cambridge University Press; 1998.
28. Zong W, Huang G- B, Chen Y. Weighted extreme learning machine for imbalance learning. Neurocomputing. 2013; 101: 229–242.
29. Younes M. The case for using digital EEG analysis in clinical sleep medicine. Sleep Science and Practice. 2017; 1(1): 2.
30. Klösch G, Kemp B, Penzel T, et al. The SIESTA project polygraphic and clinical database. IEEE Eng Med Biol Mag. 2001; 20(3): 51–57.
31. Younes M, Ostrowski M, Soiferman M, et al. Odds ratio product of sleep EEG as a continuous measure of sleep state. Sleep. 2015; 38(4): 641–654.
32. Younes M, Thompson W, Leslie C, Egan T, Giannouli E. Utility of Technologist Editing of Polysomnography Scoring Performed by

a Validated Automatic System. Ann Am Thorac Soc. 2015; 12(8): 1206–1218.
33. Younes M, Hanly PJ. Minimizing Interrater Variability in Staging Sleep by Use of Computer-Derived Features. J Clin Sleep Med. 2016; 12(10): 1347–1356.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *SLEEP* online.

ACKNOWLEDGMENTS

Thank Dr. Liyanaarachchi Lekamalage Chamara Kasun from School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, for sharing the computing resource.

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication August, 2017
Submitted in final revised form February, 2017
Accepted for publication August, 2017
Address correspondence to: M. Brandon Westover, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA. Telephone: 650-862-1154; Email: mwestover@mgh.harvard.edu

DISCLOSURE STATEMENT

MBW reports grants from NIH-NINDS (NIH-NINDS 1K23NS090900). MTB reports grants from the Department of Neurology (MGH), Milton Foundation, American Sleep Medicine Foundation, MGH-MIT Grand Challenge, and the Center for Integration of Medicine and Innovative Technology; has research contracts with MC10, Inc., and Insomnisolv, Inc.; and has consulting agreements with McKesson Health and International Flavors and Fragrances; and serves as a Medical Monitor for Pfizer, Inc; and provides expert testimony in sleep medicine. OS is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centres in Singapore Funding Initiative. The funding sources had no role in the design and conduct of the study; collection, management, analysis, or interpretation of the data; preparation, review, or approval of the manuscript.