Published in final edited form as:
J Sleep Res. 2012 February ; 21(1): 101–112. doi:10.1111/j.1365-2869.2011.00935.x.

Classification algorithms for predicting sleepiness and sleep apnea severity

Nathaniel A. Eiseman[1], **M. Brandon Westover**[1], **Joseph E. Mietus**[2], **Robert J. Thomas**[3,4], and **Matt T. Bianchi**[1,4]

[1]Neurology Department, Massachusetts General Hospital, Boston, MA, USA

[2]Division of Interdisciplinary Medicine and Biotechnology, Beth Israel Deaconess Medical Center, Boston, MA, USA

[3]Department of Medicine, Division of Pulmonary, Critical Care and Sleep, Beth Israel Deaconess Medical Center, Boston, MA, USA

[4]Division of Sleep Medicine, Harvard Medical School, Boston, MA, USA

SUMMARY

Identifying predictors of subjective sleepiness and severity of sleep apnea are important yet challenging goals in sleep medicine. Classification algorithms may provide insights, especially when large data sets are available. We analyzed polysomnography and clinical features available from the Sleep Heart Health Study. The Epworth Sleepiness Scale and the apnea–hypopnea index were the targets of three classifiers: k-nearest neighbor, naive Bayes and support vector machine algorithms. Classification was based on up to 26 features including demographics, polysomnogram, and electrocardiogram (spectrogram). Naive Bayes was best for predicting abnormal Epworth class (0–10 versus 11–24), although prediction was weak: polysomnogram features had 16.7% sensitivity and 88.8% specificity; spectrogram features had 5.3% sensitivity and 96.5% specificity. The support vector machine performed similarly to naive Bayes for predicting sleep apnea class (0–5 versus >5): 59.0% sensitivity and 74.5% specificity using clinical features and 43.4% sensitivity and 83.5% specificity using spectrographic features compared with the naive Bayes classifier, which had 57.5% sensitivity and 73.7% specificity (clinical), and 39.0% sensitivity and 82.7% specificity (spectrogram). Mutual information analysis confirmed the minimal dependency of the Epworth score on any feature, while the apnea–hypopnea index showed modest dependency on body mass index, arousal index, oxygenation and spectrogram features. Apnea classification was modestly accurate, using either clinical or spectrogram features, and showed lower sensitivity and higher specificity than common sleep apnea screening tools. Thus, clinical prediction of sleep apnea may be feasible with easily obtained demographic and electrocardiographic analysis, but the utility of the Epworth is questioned by its minimal relation to clinical, electrocardiographic, or polysomnographic features.

Correspondence, Matt T. Bianchi, Wang 7, Neurology Department, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA. Tel.: 617-724-7426; fax: 617-724-6513; mtbianchi@partners.org.

CONFLICTS OF INTEREST

Dr Bianchi, Dr Westover and Nathaniel Eiseman have no conflicts of interest to report. Dr Thomas has consulted for Total Sleep Holdings, has a patent for CO_2 adjunctive therapy for complex sleep apnea, and an ECG-based method to assess sleep stability and phenotype sleep apnea. Dr Thomas and Mr Mietus are co-inventors of the sleep spectrogram method (licensed by the BIDMC to Embla), and share patent rights and royalties.

Keywords

correlation; distribution; mutual information

INTRODUCTION

Obstructive sleep apnea (OSA) represents an under-diagnosed yet treatable risk factor for medical morbidities and daytime sleepiness (Epstein *et al.*, 2009; Malhotra and White, 2002). Although screening algorithms are available, the sensitivity and specificity of these tests render them appropriate mainly for populations with low baseline OSA risk (Bianchi, 2009). The STOP screen (snoring, tiredness, observed apnea and high blood pressure), validated in a surgical population, had a sensitivity for detecting polysom-nography (PSG)-confirmed apnea–hypopnea index (AHI) > 5 of 65.6% [confidence interval (CI): 56.4–73.9%] and a specificity of 60.0% (CI: 45.9–73.0%) (Chung et al., 2008). Adding body mass index (BMI), age, neck circumference and gender (BANG) improved the sensitivity for AHI > 5 to 83.6% (CI: 75.8–89.7), although the specificity was lower at 56.4% (CI: 42.3–69.7). If the STOP–BANG screen were applied in the surgical population used to develop it, with a approximately 70% pre-test probability of OSA (AHI >5, which was correlated with adverse surgical outcomes), a negative result would only reduce the disease probability to approximately 40% (based on the negative likelihood ratio of 0.29). The pretest probability would have to be under approximately 30% for a negative STOP–BANG screen to reduce the OSA probability to less than 10%. However, many populations, such as the surgical population just considered, have high OSA prevalence, such as patients with refractory epilepsy (33%) (Malow *et al.*, 2000), recent stroke (58%) (Bassetti *et al.*, 2006), refractory hypertension (63%) (Logan *et al.*, 2001), heart failure (35%) (Sin *et al.*, 1999) or morbid obesity undergoing bariatric surgery (80%) (Lopez *et al.*, 2008). A recent review of OSA screening tools reported pooled sensitivity of 72% (CI: 66–78%) and specificity of 61.0% (CI: 55–67%) for sleep-disorders patients, while pooled analysis of non-sleep-disorders patients revealed sensitivity of 77% (CI: 73–80%) and specificity of 53.0% (CI: 50–57%) (Abrishami *et al.*, 2010).

There is ongoing need for better predictors of sleep apnea, as well as better characterization of the relationship between apnea severity and daytime sleepiness. This goal remains particularly challenging regarding subjective endpoints such as daytime sleepiness, as several reports suggest weak or absent correlations of objective polysomnogram (PSG) parameters with the Epworth Sleepiness Scale (ESS) (Benbadis *et al.*, 1999; Chervin and Aldrich, 1999; Chervin *et al.*, 1997). For example, analysis of the large Sleep Heart Health Study database revealed a small but statistically significant relationship between categorical apnea severity (none, mild, moderate, severe) and ESS (Gottlieb *et al.*, 1999). However, even in the most severe apnea category [respiratory disturbance index (RDI) > 30], the mean ESS score (9.3) was within the normal range. Predicting sleepiness may have important policy implications, especially for management of those with alertness-sensitive occupations (Tregear *et al.*, 2009).

Two fundamental questions therefore remain unresolved: (i) can routine clinical characteristics predict apnea severity and (ii) can routine PSG features predict subjective daytime sleepiness? Because sleep apnea and daytime sleepiness are probably complex functions of many potentially interacting variables, the task of investigating predictive factors may be well suited to analysis by classification algorithms (also called ‘machine learning’ algorithms). These algorithms provide a powerful alternative to traditional regressions and correlations. Although many varieties of these algorithms exist, the unifying concept is that they can learn’ statistical patterns in a given data set (the ‘training set’) and recognize these patterns in new data (the 'testing set'). In 'supervised' learning, also called classification, the algorithm uses training set data that have already been assigned to various classes, such as patient demographic data paired with, for instance, hypertension status (classified as present or absent). The algorithms attempt to discover patterns in the feature set (e.g. demographics) associated with the provided class assignments (hypertension or not), which can then support future classification of new data. In this paper we used three supervised algorithms, the naive Bayes classifier, *k*-nearest neighbor (*k*-NN) and support vector machine (SVM), to explore whether demographic and PSG features from the Sleep Heart Health Study (SHHS) could predict the ESS, and whether routine clinical features could predict the presence of OSA (AHI > 5). In addition, we tested whether novel electrocardiogram (ECG)–spectrographic features could predict either abnormal ESS or the presence of OSA (Thomas *et al.*, 2005, 2007, 2009).

MATERIALS AND METHODS

Subjects and study design

The SHHS, a large database of home-based polysomnography (PSG) (Quan *et al.*, 1997). Category IV Institutional Review Board (BIDMC) approved use of these data, which are anonymous, and thus we did not require additional consent. The SHHS is a multi-center longitudinal study of 6441 participants drawn from several ongoing cohort studies, aged ≥40 years, designed to determine the cardiovascular consequences of sleep apnea. The baseline assessments included an overnight polysomnogram, scored using conventional rules. From this SHHS database we analyzed a subset of subjects for the current study, which did not include participants in the Strong Heart Health Study (541 subjects). Wealso excluded subjects for whom spectrogram data could not be obtained, such as excessive ECG signal dropout (< 80% of signal available for analysis), atrial fibrillation, ventricular bigeminy, demand ventricular pacing and biventricular pacing, as these conditions would interfere with single-lead ECG analysis. From the original SHHS cohort, a total of 5299 subjects were analyzed in the present analysis. As detailed in the results, the analysis was limited to subjects with complete data, $n = 4647$ (see Fig. 1).

Sleep data collection and scoring

In-home polysomnography in the SHHS was performed with 12-lead Compumedics PS (Melbourne, Australia) equipment. Manual sleep stage scoring was performed at a central location, with the following stage designations: non-rapid eye movement (NREM) Stages 1–4, REM sleep and wakefulness. Obstructive apnea was defined as an absence of airflow on the nasal cannula and a reduction in the oral thermistor signal to < 10% of baseline with continued respiratory effort, while central apneas were scored when there was no evidence of respiratory effort. Hypopnea was defined as a 30% reduction in thermistor or respiratory effort signals. The frequency (per hour of sleep) of all apneas and hypopneas associated with 4% oxygen desaturation is referred to as the AHI. The RDI used here refers to apneas and hypopneas that were associated with cortical EEG arousal.

ECG-derived sleep spectrogram

Details of the method have been published previously (Thomas *et al.*, 2005). In brief, using a continuous singlelead ECG from the PSG, we combine information from heart rate variability and ECG-derived respiration. The latter reflects amplitude variations in the QRS complex related to respiration. After filtering for outliers and cubic spline resampling at 2 Hz, the cross-spectral power and coherence of these two signals are calculated over a 1024-sample (8.5 min) window using the fast-Fourier transform applied to three overlapping 512-sample subwindows within the 1024-sample coherence window. The window is then advanced by 256 samples (2.1 min) and the computations are repeated.

For each 1024-sample window the product of the coherence and cross-spectral power is used to calculate the ratio of coherent cross-power in the low-frequency (0.01–0.1 Hz.) band to that in the high-frequency (0.1–0.4 Hz.) band. This ratio is used to classify each successive sampling window as high-frequency coupling (HFC) associated with ‘stable’ sleep or low-frequency coupling (LFC) associated with ‘unstable’ sleep. Very low-frequency coupling (VLFC) is associated with wake or REM sleep, and is calculated using the ratio of coherent cross-power in the 0–0.01 Hz band to the power in the 0.01–0.4 Hz band. A subset of LFC with especially large amplitude, called elevated-LFC (e-LFC), reflects apneic and non-apneic sleep fragmentation. It is important to note that the low-frequency component of the sleep spectrogram is not equivalent to the low-frequency component of the HRV spectrum, but more specifically represents relatively low-frequency respiratory coupled oscillations in heart rate. Thus, the frequency bands of the ECG-derived sleep spectrogram are distinct from the standard HRV bands, although an overlap exists.

Classification

We pre-defined a series of $n = 27$ features of interest, including two that would be used as targets for classification: the ESS and the AHI. These features included routinely available clinical information (age, sex, race, blood pressure, presence of diabetes, hypertension, coronary disease or angina, BMI, ESS), routine PSG features and the ECG spectrogram calculated from the sleep study ECG channel.

The k-NN algorithm focuses on local patterns in the feature space defined by these features for each subject. The value of k determines how local’ the algorithm will restrict its search for patterns in the feature space. k-NN can thus capture local patterns or clusters in the data set. As k approaches the number of subjects in the data set, classification occurs according to the most prevalent class in the entire set.

SVM classifiers use one of several types of kernel functions to specify a hyperplane in multi-dimensional space in order to perform classification (see http://www.support-vector-machines.org/for review). We implemented the most commonly used SVM, based on the radial basis function kernel. We manually tested a range of values for the C, gamma and epsilon parameters. The C parameter is a penalty term: small C values tend to under-fit the data (increased errors), while high C values tend towards over-fitting, with asymptotic approach to the ‘hard margin’ condition as C approaches infinity. The gamma parameter refers to the smoothness of the boundaries of the hyperplane: higher gamma values allow more irregular boundaries, which corresponds to an increased risk of over-fitting. The epsilon parameter represents the ‘insensitivity zone’, or tolerance for classification errors. Higher epsilon values reduce the accuracy requirement during training, and decrease the number of support vectors in the classifier (important for avoiding overfitting). For AHI classification using combined clinical, ECG and PSG features, we tested manually a range of parameter values: epsilon (0–0.5), C (0.1–100) and gamma (0.01–100). When C and gamma were both 1 or higher, performance was poor regardless of epsilon (LR values very close to 1). When gamma was 0.01–0.3, C was 0.1–1 and epsilon was 0–0.2, the performance varied smoothly over a sensitivity range of approximately 35–65% and specificity of 65–85%. For ESS, this range of parameters yielded extremely poor performance (sensitivity approximately 0%, specificity approximately 100%, reflecting the higher prevalence of normal ESS class).

In our analysis, optimal classification performance involved a number of support vectors that equaled the number of subjects (i.e. 4647), which suggests that class discrimination was difficult. The time required to train the SVM classifier ranged from 3 minutes to >2 hour for a single set of parameters (on a Dell Core 2 Duo laptop), much slower than the naive Bayes classifier.

The naive Bayes classifier assumes that any given feature value (such as BMI) for a particular class (such as AHI > 5) is independent of the other feature values for that class. It employs a maximum likelihood method of parameter estimation. By assuming independence of each feature (‘naive’), the N-dimensional problem is effectively reduced to the computationally more simple circumstance of N one-dimensional problems. Empirically, the independence assumption does not often affect classification accuracy as much as one would expect; in fact, naive Bayes classifiers perform unexpectedly well on many real-world classification problems despite the often unrealistic assumption of feature independence in classification (Zhang, 2004). To demonstrate further that the independence assumption did not compromise our results, the data set was considered large enough to also perform non-naive Bayes classification of AHI and ESS using the ECG features (and possibly large enough for the clinical features). This method, in fact, reduced to the independence assumption for the ECG features, as well as the nine clinical features (we did not attempt this for all features, due to insufficient subjects). In other words, there was no advantage to considering added information in combining features in this data set (data not shown).

Classification algorithms were implemented using the freely available software RapidMiner (http://rapid-i.com/content/view/181/190/lang,en/). For retrospective analysis such as this, the classification algorithms utilized a validation method known as K-fold cross-validation. The training set is defined here as the subset of the 5299 subjects described above with complete data ($n = 4647$). This training set is then divided into K equal subsets (we used $K = 20$), then K–1 of the subsets are used to train the algorithm, while the remaining one subset is used to apply the learned classification. This process is repeated K times, such that each subset is classified once by the algorithm trained on the remaining data; that is, all subjects participate in training sets multiple times and in the classification test set once. The most accurate validation results are obtained using the method of ‘stratified sampling’ to obtain the subsets, whereby the distribution of classes in the subsets is as similar as possible to that of the entire training set. The results are aggregated into a classification accuracy table (known as a ‘confusion matrix’), which shows correctly and incorrectly classified subjects, from which sensitivity, specificity and predictive values can be calculated.

RESULTS

The distribution of clinical features from the SHHS database

The feature set we considered from the SHHS database included routinely available clinical features [age, sex, race, BMI, systolic and diastolic blood pressure (SBP and DBP), coronary artery disease (CAD; defined by a history of angina or myocardial infarction), diabetes (DM) and ESS] and PSG features [AHI, RDI, total sleep time (TST), sleep efficiency, percentage of stage N1, N2, N3 and REM sleep, arousal index (total, as well as stage specific: NREM versus REM sleep), percentage of the night below 90% oxygen saturation, and oxygen nadir in REM and NREM sleep] (Table 1). In addition, we considered novel ECG-spectrographic features that characterize sleep architecture and sleep-disordered breathing by the dominant frequency of cardiopulmonary coupling, consisting of a combination of autonomic heart rate variability and respiration-related changes in *R*-wave amplitude (Thomas *et al.*, 2005, 2009, 2010). The ECG-spectrogram allows sleep to be categorized according to the amount of time spent in states of high-frequency cardiopulmonary coupling (HFC; associated with stable respiratory rate and tidal volumes in NREM sleep), LFC (associated with fluctuations in respiratory rate and tidal volumes), e-LFC (associated with apneas and hypopneas) and VLFC (associated with fluctuations characteristic of wake and of REM sleep) (Table 1). We restricted analysis to the subjects in the SHHS who had an adequate ECG signal (see Materials and methods) in order to use the ECG-spectrographic features, as reported previously (Thomas *et al.*, 2009). Thus, a total of

27 features were considered, two of which were chosen as targets for algorithm classification: AHI and ESS.

We divided the features into three categories (clinical, PSG and spectrographic) to represent types of data that might be considered when making clinical predictions. For example, the problem of predicting sleep apnea is mainly relevant when considering non-PSG features, as obtaining the PSG itself includes routine measurement of AHI, so making an AHI prediction based upon, for instance, sleep stages, would be of mainly academic interest. However, being able to predict AHI based on purely clinical features or a simple single-lead ECG would have potential practical utility in screening or risk stratification. Similarly, predicting ESS based on physiology is of interest for mechanistic reasons; however, predicting it based on clinical features is less useful because the ESS is itself obtained easily in routine clinical contact.

Figure 1 shows the distribution of values for routine clinical features. Binned histograms of the continuous variables demonstrated non-Gaussian distributions in each case within this data set. For each continuous variable, we found statistically significant deviation from normality in all cases by two tests (the KS normality test and the D'Agostino and Pearson normality test). A minority of subjects contained missing values for at least one of the 27 features. Thus, we restricted the data set further for classification to those subjects with complete data. In order to assess whether subjects with missing data differed systematically from those with complete data ($n = 4647$), each histogram is overlaid with the distribution of subjects in the 'missing data' subset ($n = 652$). Subjects with missing data had similar distributions to those with complete data in each case, such that removing this subset from the classification approach is unlikely to confound the results.

Correlation analysis of clinical features in the SHHS database

We calculated non-parametric correlations (Spearman's rank method) between individual clinical, PSG and ECG features (discrete/continuous metrics only; $n = 22$) and the two variables of interest, ESS and AHI. Small but significant correlations were found between most features and the ESS (Fig. 2a) and AHI (Fig. 2b). The strongest correlations with ESS were small, and the only ones with an r-value at or stronger than 0.1 (or −0.1) were the AHI (0.13), the RDI (0.1), the BMI (0.1), the % time with <90% oxygen saturation (0.1), the REM oxygen nadir (−0.1) and the NREM oxygen nadir (−0.11). The AHI, in contrast, was correlated more strongly with several features, and all the features had r-values stronger than 0.1 (or −0.1) (Fig. 2b). Scatterplots are shown for two commonly associated feature-pairs in Fig. 2c (ESS versus AHI) and Fig. 2d (AHI versus BMI). These plots illustrate the variability in the data, consistent with the modest correlations, and suggest the potential utility of more sophisticated classification methods to capture patterns and relationships between features and the endpoints of ESS and AHI.

Naive Bayes performance in sleepiness classification

A naive Bayes classifier assigns the test data points (subjects) in question based on the probability of each feature of that subject occurring in a given class (in this study, defined by AHI and ESS score). In other words, each feature may be considered to have a sensitivity and specificity with regard to a given class membership. Thus, the combination of sensitivity, specificity and prior probability of class membership is used by the algorithm according to Bayes' theorem. The 'naive' aspect refers to the fact that the algorithm assumes that the features are independent of each other with regard to class association. This assumption may not reflect the clinical reality of interactions among patient features; we address this below, by comparing the performance of a non-naive Bayes classifier algorithm.

We first attempted to classify ESS using the naive Bayes classifier, based on either PSG features or ECG spectrographic features. ESS values were dichotomized into normal (0–10) or abnormal (11–24) according to typical clinical criteria. Algorithm performance is shown in the form of a 'confusion matrix', which is of similar structure to the familiar dichotomous 2×2 box illustrating sensitivity, specificity and predictive value of diagnostic test performance. Sensitivity and specificity values determine the positive and negative likelihood ratios, according to the following equations: $LR^{(+)}$ = sensitivity/(100 – specificity) and $LR^{(-)}$ = (100 – sensitivity)/specificity.

The classification accuracy for ESS class was poor regardless of training on PSG or spectrogram data (Fig. 3-a,b). Using PSG data, the sensitivity for detecting abnormal ESS was only 16.7%, while the specificity was 88.8%, corresponding to $LR^{(+)}$ of 1.49 and $LR^{(-)}$ of 0.94. Using the ECG data, sensitivity was lower at 5.3% and specificity higher at 96.5%, but the LR values were still poor, with $LR^{(+)}$ of 1.51 and $LR^{(-)}$ of 0.98. The closer the LR values are to 1, the smaller will be the change in disease probability after obtaining the test result, according to Bayes' theorem. Finally, we tested whether all available information (26 features) would improve the ESS classification (Fig. 3c), but the results were similar to those obtained with clinical features only (Fig. 3a).

Interpreting the confusion matrices requires consideration of the proportion of subjects in each class—that is, the prior probability or prevalence. In the SHHS population used for this analysis, the prevalence of abnormal ESS was approximately 25%, and thus the PPV of the algorithm (approximately 34–36%) represents only a small improvement over this prevalence value. The NPV was nearly identical to the prevalence of normal ESS, as expected when the sensitivity is so low and the $LR^{(-)}$ is so close to 1. In other words, these LR values indicated that the classification algorithm, viewed as a diagnostic test, yields little information beyond that contained in the prior probabilities. The classification performance was only marginally better if we markedly shifted the cutoff of abnormal ESS (for example, using 0–1 as normal, or using 0–19 as normal), suggesting that the poor performance was not simply attributable to the clinical definition of normal ESS as 0–10 (data not shown). Finally, to assess the possibility that features were not independent (as assumed by the algorithm), we tested the non-naive analog of this classifier for ECG and clinical features. The classification method, however, reduced to the naive case, indicating that the learning algorithm could not find statistical evidence of dependence between any of the features. Although this does not mean that dependencies do not exist (for example, there are known dependencies between AHI and RDI), it suggests that combining features does not improve significantly the performance of the classification algorithm.

The classification performance was worse using the *k*-NN algorithm, with tested k values of 1, 3, 5 and 10 (data not shown). The poor classification accuracy for ESS with both algorithms could be either because the features truly do not predict sleepiness class, or because the ESS is a poor marker of sleepiness, or both. For example, the eight questions of the ESS are equally weighted, although it is likely that falling asleep while driving is a more substantial indicator of sleepiness than falling asleep after laying down in the afternoon explicitly to rest.

Naive Bayes performance in classifying apnea severity

We next turned to prediction of an objectively defined metric, the AHI, using clinical features and ECG features. We dichotomized all subjects into normal (0–5) or abnormal (>5), based on the typical clinical criteria for diagnosing OSA. AHI classification was performed based on clinical features (Fig. 3c) or ECG features (Fig. 3d). Using clinical features, the sensitivity for AHI >5 was 57.5% and specificity was 73.7%, corresponding to

$LR^{(+)}$ 2.19 and $LR^{(-)}$ 0.58. Using ECG features, the sensitivity was lower at 39.0%, with a higher specificity of 82.7%, corresponding to $LR^{(+)}$ 2.25 and $LR^{(-)}$ 0.74.

Although these results demonstrated improved prediction of the more objective endpoint of AHI compared to the subjective ESS, the LR values are still close to 1, and thus only modestly adjust disease probability. For example, in the subjects studied here, the prevalence of AHI >5 was approximately 45%, such that the PPV using either clinical-or ECG-based classifiers was approximately 65%. The NPV was approximately 68% based on clinical features and approximately 62% based on ECG features.

Classification was not improved substantially when four groups of AHI were considered (0–5, 5–15, 15–30 and >30), including individual cutoffs such as <30 versus >30 (data not shown). As in the case of ESS classification, the *k*-NN algorithm performance for AHI classification was worse, for *k* values of 1, 3, 5 and 10 (data not shown).

Finally, we performed AHI classification using a combination of clinical, spectrographic and PSG features unrelated to sleep-disordered breathing. We excluded PSG metrics of RDI, low O_2 values and arousal indices because they are tied intimately to the actual calculation of AHI, and may thus elevate falsely the classification accuracy for trivial reasons. Using these combined data, the sensitivity for AHI > 5 was 56.0% and specificity was 77.4%, essentially unchanged from the classification based on clinical features (Fig. 3f).

Predicting sleepiness and sleep apnea with an SVM classifier

We next turned to the SVM classification technique. SVM considers the distribution of class features in multi-dimensional space and designates a ‘hyperplane’ that allows the best feature-based separation of the classes of interest. We used a radial basis function kernel, which is very flexible in the consideration of non-linear feature relationships. ESS classification was poor with this method, and across a range of parameter combinations the classification defaulted to the prior probability: that is, the algorithm classified all subjects into the normal 0–10 class, which was the most prevalent.

Classification of AHI, in contrast, was much better, and performed similarly to the naive Bayes classifier (Fig. 4) for clinical and spectrographic features. Using clinical features, the sensitivity was 59.0% and the specificity was 74.5%, with PPV 65.4% and NPV of 69.0% (Fig. 4a). The corresponding $LR^{(+)}$ was 2.3, and the $LR^{(-)}$ was 0.55. Using spectrographic features, the sensitivity was 43.4% and the specificity was 83.5%, with PPV 68.3% and NPV of 64.4% (Fig. 4b). The corresponding $LR^{(+)}$ was 2.6, and the $LR^{(-)}$ was 0.68. Combining these features with non-respiration PSG features yielded sensitivity of 62.3% and specificity 78.3%, with PPV 70.1% and NPV of 71.7%. The corresponding $LR^{(+)}$ was 2.9 and the $LR^{(-)}$ was 0.48. The combined data performed slightly better than the naive Bayes classifier.

The weights assigned to the features used in the ‘combined’ data set are shown in Fig. 4d. The relative importance of each feature is in general agreement with clinical expectation and the non-parametric correlation data shown in Fig. 2. For example, the e-LFC, BMI and LFC features were the strongest in the algorithm. We again note that although the ‘combined’ data improved the classification, the aim of classification is to perform well without the need for PSG features (which already include the AHI in routine practice), and thus we suggest that the most practical results are those that classify based on easily obtained clinical or ECG features.

Mutual information approach to correlating clinical features with ESS and AHI

Finally, we undertook an information theoretical approach to quantifying the relationship of various features with the ESS and the AHI. Mutual information is a powerful tool in this regard because it captures how much statistical information one variable can provide about another. For the SHHS data, we thus used mutual information to determine the relationship between clinical, PSG and ECG features and the ESS or AHI. Mutual information is not limited to linear relationships as in the Pearson's correlation, or to monotonic relationships as in the non-parametric Spearman's rank correlation. Instead, it captures any relationship (sometimes referred to as dependency) between the variables—without needing to know or specify what the relationship is. Because this calculation depends to some extent upon the number of bins used to categorize the ESS and AHI, we normalized the mutual information value (which, like entropy, is in units of bits) to the entropy of the ESS and the AHI distributions themselves. In this way, values approaching zero indicate little or no shared information or dependency, while values approaching 1 indicate a high or exact degree of dependency (of any kind) between the two variables.

Figure 5a shows the normalized mutual information between ESS and multiple discrete/ continuous features. In every case the value was close to zero, and always < 0.05, indicating very little shared dependency, consistent with the poor performance of the classification algorithms for predicting ESS. Note that even the AHI has a nearly zero value, emphasizing the exceedingly small statistical dependency between the AHI and ESS. Figure 5b shows the normalized mutual information between the features and the AHI. As expected, there were several features that showed some degree of dependency. For example, the arousal index (whether total, in NREM or in REM sleep) showed a small relationship with AHI. The RDI and the oxygenation metrics also showed dependency, as expected, as the RDI depends in part on the AHI, which includes oxygen values. Finally, the ECG features showed a relationship with AHI. This is also expected as the ECG-spectrogram (in particular e-LFC) has been associated with apnea severity (Thomas *et al.*, 2009).

DISCUSSION

This study used the large standardized SHHS database to determine whether classification algorithms could predict two relevant endpoints—AHI > 5 and ESS > 10—from collections of clinical, PSG or cardiorespiratory /autonomic features. The following conclusions can be drawn from this analysis: (i) performance in predicting ESS was poor, regardless of which features were used or which ESS threshold was considered; (ii) performance in predicting AHI > 5 was notably better, but still only modest sensitivity and specificity values were obtained; (iii) SVM performed slightly better than the naive Bayes classifier for predicting AHI class (but not ESS class) at the expense of the need for parameter searching and larger training times; and (iv) mutual information analysis provided a basis for the discrepancy in prediction accuracy, because ESS values showed essentially no dependency on any features.

Prediction of the Epworth Sleepiness Scale

The ESS requires patients to reflect on the chances of dozing, in general (not for any particular time frame), for each of eight circumstances. Several studies have suggested little or no correlation of ESS scores with AHI values or with Multiple Sleep Latency Test (MSLT) values (Chervin, 2000; Chervin and Aldrich, 1999; Chervin *et al.*, 1997; Gottlieb *et al.*, 1999), and alternative measures to predict sleepiness have been proposed (Chervin and Aldrich, 1998; Chervin *et al.*, 2005), as well as alternative analysis techniques of the ESS (Smith *et al.*, 2008). The inverse question, of whether sleepiness captured by the ESS score predicts apnea severity (Gottlieb *et al.*, 1999), is also interesting as this subjective complaint may be among early clues to the clinical suspicion of sleep-disordered breathing. Although the ESS score was used in one clinical predictor of OSA (Santaolalla Montoya *et al.*, 2007), it was not found to be of value in a neural network clinical predictor (Kirby *et al.*, 1999), and is not used in several other clinical prediction algorithms (Roche *et al.*, 2002; Rowley *et al.*, 2000; Young *et al.*, 2002). Several considerations may explain the weak classification performance observed here as well as in prior studies seeking association with PSG and MSLT findings. The composite measure of equally weighted circumstances in the scale may obscure otherwise useful information contained in individual responses. Also, the subjective sense of sleepiness may be impacted by comorbid illness, rate of development of condition(s) causing sleepiness, tolerance to challenges causing sleepiness (such as OSA) and /or countermeasures such as caffeine. Moreover, the cutoff value of >10 for abnormal may not be generally applicable. We addressed this final consideration by choosing various class cutoff values, such as isolating the extreme values which may be more specific. However, neither extreme values nor breaking the scores into quartiles provided any improvement. Together with the mutual information analysis, the results suggest that the ESS score has little dependency on clinical, PSG or autonomic features, and thus is therefore inherently difficult to predict. Considering the weak relation of ESS to the objective measure of sleepiness provided by the MSLT, the results suggest the need for further efforts to improve quantification of subjective sleepiness.

Prediction of the AHI

Predicting the presence and /or severity of sleep apnea is of great interest from the general preventative care screening setting, to optimal utilization of laboratory PSG resources, to specialized settings such as post-operative care units. Screening questionnaires such as the Berlin questionnaire, the Wisconsin Sleep questionnaire and the STOP–BANG questionnaire have the advantages of being straightforward, brief and inexpensive to administer in the ambulatory setting—however, they have only modest sensitivity and specificity (Abrishami *et al.*, 2010). More sophisticated methods have been used to establish predictors of OSA based on statistical analysis of various clinical features. The highest sensitivity and specificity values were obtained in a retrospective neural network classifier trained on various clinical features (Kirby *et al.*, 1999). The prevalence of OSA (AHI>10) in this group of 405 patients was nearly 70%, and the network performed well with sensitivity of 99% and specificity of 80%. Logistic models based on routine clinical features fared less well, identifying AHI >10 in 370 patients referred for OSA with 76–96% sensitivity and 13–54% specificity (prevalence of OSA in that cohort was 67%) (Rowley *et al.*, 2000). Identifying those with AHI >20 had lower sensitivity and higher specificity, and the authors suggested that the prediction rules might be useful for stratifying patients for split-night studies, but were not accurate enough for general screening use. Multiple logistic regression of clinical features available in the SHHS suggested several independent predictors of AHI>15, including male sex, age, BMI, neck girth, snoring and frequency of reported nocturnal respiratory pauses (Young *et al.*, 2002), but no explicit prediction model with sensitivity and specificity for OSA was reported. Roche *et al.* (2002) developed a multiple linear regression analysis with reasonable predictive value in their training cohort, but the model performed poorly on a validation cohort from their center, emphasizing the importance of validating prediction models prospectively.

Regardless of the method used as a screening test or predictor of OSA, one prevailing challenge involves the apparently high prevalence of OSA in various clinical populations as described above. When the pre-test probability of OSA is high, any screening test must have fairly small $\mathrm{LR}^{(-)}$ values in order trust that a negative test result is not simply a false negative. The 2007 American Academy of Sleep Medicine guidelines for home sleep monitoring suggest limiting use to those with high pre-test probability of disease (Collop *et al.*, 2007)—in this population the risk of false negatives should warrant caution. It is worth also pointing out that a high sensitivity, considered typically as critical for good 'rule-out' power, is not sufficient—the specificity is also critical. For example, any time the sensitivity and specificity percentages add to 100%, the positive and negative LR values will be 1; that is, no change in probability with either test result. In our ESS prediction, the high specificity is therefore tempered by the extremely low sensitivity, and thus the $LR^{(+)}$ remained low, such that a positive classification provided little adjustment in the probability of abnormal ESS class.

Although the AHI does not suffer from the subjective complexities of the ESS, it is sensitive to a variety of factors that may change from night to night within an individual, and thus the AHI values obtained from single PSG assessments in the SHHS may not reflect each subject's 'true' apnea index. For example, body position, amount of REM sleep, presence of intermittent nasal congestion, sleep drive on the particular night of study or other stochastic fluctuations in apnea severity contribute to this variability, such that a single night of recording may not constitute an adequate sample (Levendowski *et al.*, 2009). We did not include other morphometric values that may be associated with severity.

Finally, it is worth mentioning that, due to the nature of the internal cross-validation process used here, it is difficult to predict how machine learning algorithms might perform in other data sets. It would be interesting to apply similar classification algorithm approaches to other data sets, as the SHHS may not be representative of the general population (Lind *et al.*, 2003). Different populations may be more amenable to classification if, for example, they contain less clinical heterogeneity. Also, it would be interesting to utilize different endpoints for sleepiness (such as multiple sleep latency or maintenance of wakefulness testing), and in the case of apnea severity to measure this on repeated nights, given that this measurement itself contains some variance not captured by the single night assessments in this data set.

Acknowledgments

The authors thank Scott McKinney, Drs Sydney Cash, Catherine Chu-Shore and Elizabeth Klerman for valuable discussions. We thank Alex Chan and Melissa St Hilaire for advice on support vector machine classification. Dr Bianchi receives funding from the Department of Neurology, Massachusetts General Hospital and the Clinical Investigator Training Program: Harvard /MIT Health Sciences and Technology—Beth Israel Deaconess Medical Center, in collaboration with Pfizer, Inc. and Merck & Co. This funding source had no role in the design, interpretation or publication of this study. This paper represents the work of the authors. This manuscript was not prepared in collaboration with investigators of the Sleep Heart Health Study and does not necessarily reflect the opinions or views of the Sleep Heart Health Study or the NHLBI.

REFERENCES

Abrishami A, Khajehdehi A, Chung F. A systematic review of screening questionnaires for obstructive sleep apnea. Can. J. Anaesth. 2010; 57:423–438. [PubMed: 20143278]

Bassetti CL, Milanova M, Gugger M. Sleep-disordered breathing and acute ischemic stroke: diagnosis, risk factors, treatment, evolution, and long-term clinical outcome. Stroke. 2006; 37:967–972. [PubMed: 16543515]

Benbadis SR, Mascha E, Perry MC, Wolgamuth BR, Smolley LA, Dinner DS. Association between the Epworth sleepiness scale and the multiple sleep latency test in a clinical population. Ann. Intern. Med. 1999; 130:289–292. [PubMed: 10068387]

Bianchi MT. Screening for obstructive sleep apnea: Bayes weighs in. Open Sleep J. 2009; 2:56–59.

Chervin RD. The multiple sleep latency test and Epworth sleepiness scale in the assessment of daytime sleepiness. J. Sleep Res. 2000; 9:399–401. [PubMed: 11123526]

Chervin RD, Aldrich MS. Characteristics of apneas and hypopneas during sleep and relation to excessive daytime sleepiness. Sleep. 1998; 21:799–806. [PubMed: 9871942]

Chervin RD, Aldrich MS. The Epworth Sleepiness Scale may not reflect objective measures of sleepiness or sleep apnea. Neurology. 1999; 52:125–131. [PubMed: 9921859]

Chervin RD, Aldrich MS, Pickett R, Guilleminault C. Comparison of the results of the epworth sleepiness scale and the multiple sleep latency test. J. Psychosom. Res. 1997; 42:145–155. [PubMed: 9076642]

Chervin RD, Burns JW, Ruzicka DL. Electroencephalo-graphic changes during respiratory cycles predict sleepiness in sleep apnea. Am. J. Respir. Crit. Care Med. 2005; 171:652–658. [PubMed: 15591467]

Chung F, Yegneswaran B, Liao P, et al. STOP questionnaire: a tool to screen patients for obstructive sleep apnea. Anesthesiology. 2008; 108:812–821. [PubMed: 18431116]

Collop NA, Anderson WM, Boehlecke B, et al. Clinical guidelines for the use of unattended portable monitors in the diagnosis of obstructive sleep apnea in adult patients. Portable Monitoring Task Force of the American Academy of Sleep Medicine. J. Clin. Sleep Med. 2007; 3:737–747. [PubMed: 18198809]

Epstein LJ, Kristo D, Strollo PJ Jr, et al. Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. J. Clin. Sleep Med. 2009; 5:263–276. [PubMed: 19960649]

Gottlieb DJ, Whitney CW, Bonekat WH, et al. Relation of sleepiness to respiratory disturbance index: the Sleep Heart Health Study. Am. J. Respir. Crit. Care Med. 1999; 159:502–507. [PubMed: 9927364]

Kirby SD, Eng P, Danter W, et al. Neural network prediction of obstructive sleep apnea from clinical criteria. Chest. 1999; 116:409–415. [PubMed: 10453870]

Levendowski DJ, Zack N, Rao S, et al. Assessment of the test- retest reliability of laboratory polysomnography. Sleep Breath. 2009; 13:163–167. [PubMed: 18766393]

Lind BK, Goodwin JL, Hill JG, Ali T, Redline S, Quan SF. Recruitment of healthy adults into a study of overnight sleep monitoring in the home: experience of the Sleep Heart Health Study. Sleep Breath. 2003; 7:13–24. [PubMed: 12712393]

Logan AG, Perlikowski SM, Mente A, et al. High prevalence of unrecognized sleep apnoea in drug-resistant hypertension. J. Hypertens. 2001; 19:2271–2277. [PubMed: 11725173]

Lopez PP, Stefan B, Schulman CI, Byers PM. Prevalence of sleep apnea in morbidly obese patients who presented for weight loss surgery evaluation: more evidence for routine screening for obstructive sleep apnea before weight loss surgery. Am. Surg. 2008; 74:834–838. [PubMed: 18807673]

Malhotra A, White DP. Obstructive sleep apnoea. Lancet. 2002; 360:237–245. [PubMed: 12133673]

Malow BA, Levy K, Maturen K, Bowes R. Obstructive sleep apnea is common in medically refractory epilepsy patients. Neurology. 2000; 55:1002–1007. [PubMed: 11061259]

Quan SF, Howard BV, Iber C, et al. The Sleep Heart Health Study: design, rationale, and methods. Sleep. 1997; 20:1077–1085. [PubMed: 9493915]

Roche N, Herer B, Roig C, Huchon G. Prospective testing of two models based on clinical and oximetric variables for prediction of obstructive sleep apnea. Chest. 2002; 121:747–752. [PubMed: 11888955]

Rowley JA, Aboussouan LS, Badr MS. The use of clinical prediction formulas in the evaluation of obstructive sleep apnea. Sleep. 2000; 23:929–938. [PubMed: 11083602]

Santaolalla Montoya F, Iriondo Bedialauneta JR, Aguirre Larra-coechea U, Martinez Ibarguen A, Sanchez Del Rey A, Sanchez Fernandez JM. The predictive value of clinical and epidemiological parameters in the identification of patients with obstructive sleep apnoea (OSA): a clinical prediction algorithm in the evaluation of OSA. Eur. Arch. Otorhinolaryngol. 2007; 264:637–643. [PubMed: 17256124]

Sin DD, Fitzgerald F, Parker JD, Newton G, Floras JS, Bradley TD. Risk factors for central and obstructive sleep apnea in 450 men and women with congestive heart failure. Am. J. Respir. Crit. Care Med. 1999; 160:1101–1106. [PubMed: 10508793]

Smith SS, Oei TP, Douglas JA, Brown I, Jorgensen G, Andrews J. Confirmatory factor analysis of the Epworth Sleepiness Scale (ESS) in patients with obstructive sleep apnoea. Sleep Med. 2008; 9:739–744. [PubMed: 17921053]

Thomas RJ, Mietus JE, Peng CK, Goldberger AL. An electrocardiogram-based technique to assess cardiopulmonary coupling during sleep. Sleep. 2005; 28:1151–1161. [PubMed: 16268385]

Thomas RJ, Mietus JE, Peng CK, et al. Differentiating obstructive from central and complex sleep apnea using an automated electrocardiogram-based method. Sleep. 2007; 30:1756–1769. [PubMed: 18246985]

Thomas RJ, Weiss MD, Mietus JE, Peng CK, Goldberger AL, Gottlieb DJ. Prevalent hypertension and stroke in the Sleep Heart Health Study: association with an ECG-derived spectrographic marker of cardiopulmonary coupling. Sleep. 2009; 32:897–904. [PubMed: 19639752]

Thomas RJ, Mietus JE, Peng CK, Goldberger AL, Crofford LJ, Chervin RD. Impaired sleep quality in fibromyalgia: detection and quantification with ECG-based cardiopulmonary coupling spectrograms. Sleep Med. 2010; 11:497–498. [PubMed: 20015685]

Tregear S, Reston J, Schoelles K, Phillips B. Obstructive sleep apnea and risk of motor vehicle crash: systematic review and meta-analysis. J. Clin. Sleep Med. 2009; 5:573–581. [PubMed: 20465027]

Young T, Shahar E, Nieto FJ, et al. Predictors of sleep-disordered breathing in community-dwelling adults: the Sleep Heart Health Study. Arch. Intern. Med. 2002; 162:893–900. [PubMed: 11966340]

Zhang, H. The optimality of naive bayes; Proceedings of International Florida Artificial Intelligence Researh Society Conference; 17–19 May 2004;

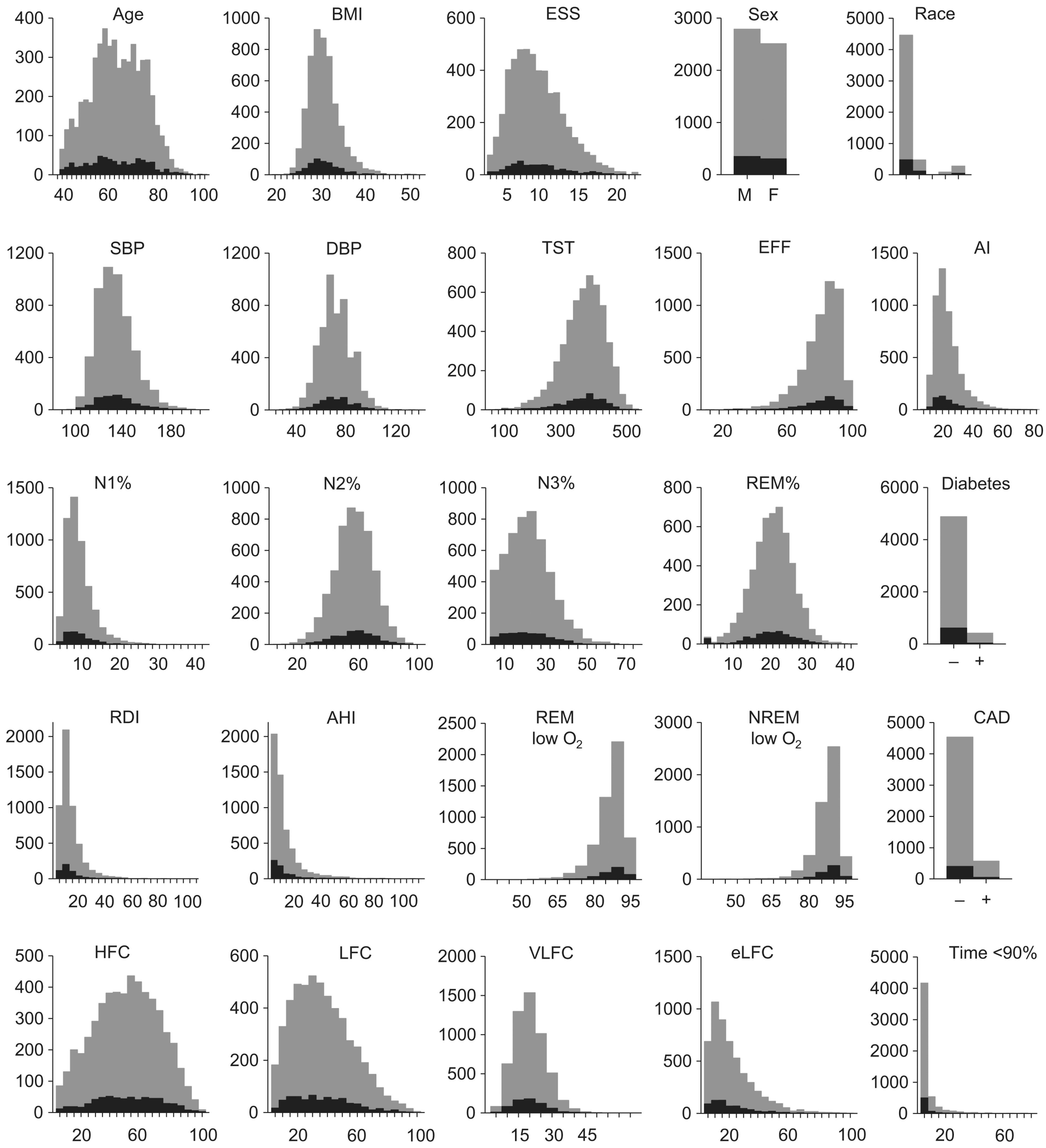


Figure 1.

Distribution of clinical features in the Sleep Heart Health Study (SHHS) subjects. Histograms of various clinical features are shown in overlapped bars, where dark grey represents subjects with at least one missing data point, and light grey representing subjects with complete data. We studied 27 features in total. Twenty-five are shown in this figure; arousal index (AI) in non-rapid eye movement (NREM) and REM are not shown, but demonstrated similar complete versus missing distributions. See Table 1 for description of features.

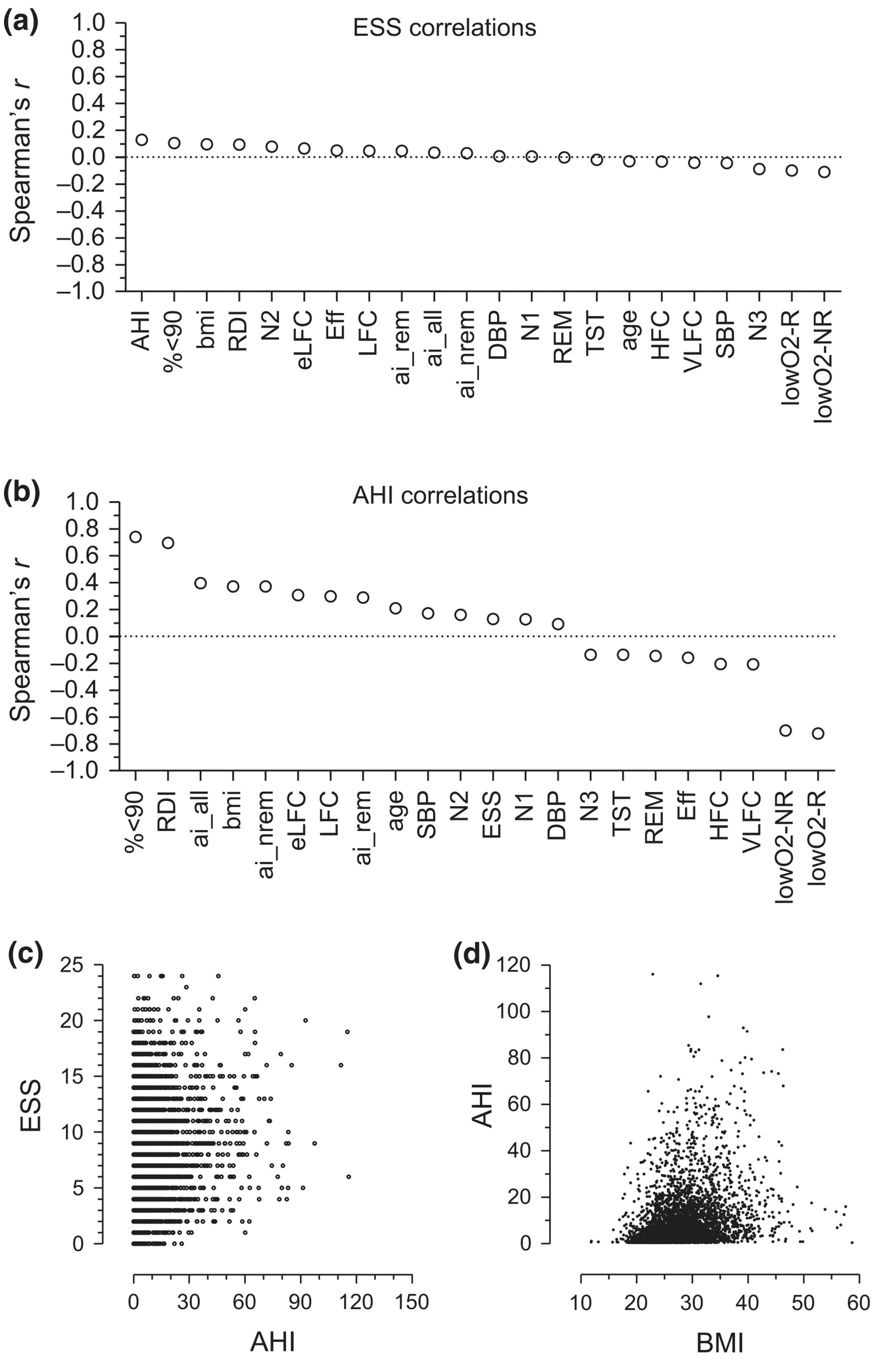
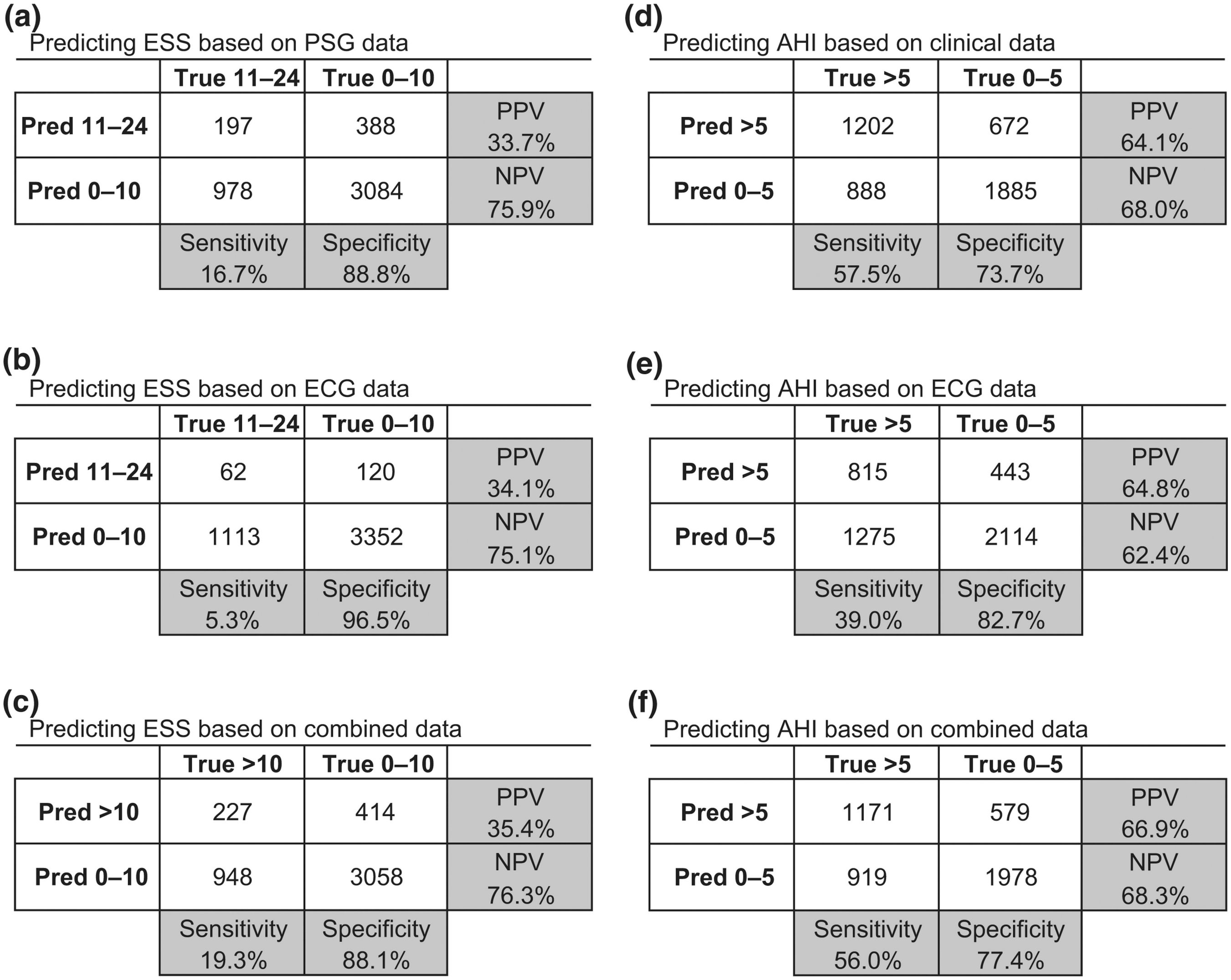


Figure 2.

Spearman's correlation of clinical features with Epworth Sleepiness Scale (ESS) and apnea–hypopnea index (AHI). *R*-values for correlation between ESS (a) and AHI (b) are shown for the listed clinical features. Scatterplots are shown for representative pairs: ESS versus AHI (c) and AHI versus body mass index (BMI) (d).



(a)

Predicting ESS based on PSG data

	True 11–24	True 0–10	
Pred 11–24	197	388	PPV 33.7%
Pred 0–10	978	3084	NPV 75.9%
	Sensitivity 16.7%	Specificity 88.8%	

(b)

Predicting ESS based on ECG data

	True 11–24	True 0–10	
Pred 11–24	62	120	PPV 34.1%
Pred 0–10	1113	3352	NPV 75.1%
	Sensitivity 5.3%	Specificity 96.5%	

(c)

Predicting ESS based on combined data

	True >10	True 0–10	
Pred >10	227	414	PPV 35.4%
Pred 0–10	948	3058	NPV 76.3%
	Sensitivity 19.3%	Specificity 88.1%	

(d)

Predicting AHI based on clinical data

	True >5	True 0–5	
Pred >5	1202	672	PPV 64.1%
Pred 0–5	888	1885	NPV 68.0%
	Sensitivity 57.5%	Specificity 73.7%	

(e)

Predicting AHI based on ECG data

	True >5	True 0–5	
Pred >5	815	443	PPV 64.8%
Pred 0–5	1275	2114	NPV 62.4%
	Sensitivity 39.0%	Specificity 82.7%	

(f)

Predicting AHI based on combined data

	True >5	True 0–5	
Pred >5	1171	579	PPV 66.9%
Pred 0–5	919	1978	NPV 68.3%
	Sensitivity 56.0%	Specificity 77.4%	

Figure 3.

Performance of the naive Bayes classifier in predicting Epworth Sleepiness Scale (ESS) and apnea–hypopnea index (AHI) classes. The prediction power of the naive Bayes classifier algorithm is shown for ESS based on polysomnogram (PSG) features (a), electrocardiogram (ECG) features (b) or a combination of all 26 available features (c). The prediction power of the naive Bayes classifier algorithm is shown for AHI based on clinical features (d), ECG features (e) or a combination of features (f). For (f), the combination included clinical, ECG and only those PSG values not related to apnea index [excluded arousal, respiratory disturbance index (RDI) and oxygen metrics].

(a) Predicting AHI based on clinical data

	True >5	True 0–5	
Pred >5	1232	652	PPV 65.4%
Pred 0–5	858	1905	NPV 69.0%
	Sensitivity 59.0%	Specificity 74.5%	

(b) Predicting AHI based on ECG data

	True >5	True 0–5	
Pred >5	907	421	PPV 68.3%
Pred 0–5	1183	2136	NPV 64.4%
	Sensitivity 43.4%	Specificity 83.5%	

(c) Predicting AHI based on combined data

	True >5	True 0–5	
Pred >5	1301	554	PPV 70.1%
Pred 0–5	789	2003	NPV 71.7%
	Sensitivity 62.3%	Specificity 78.3%	

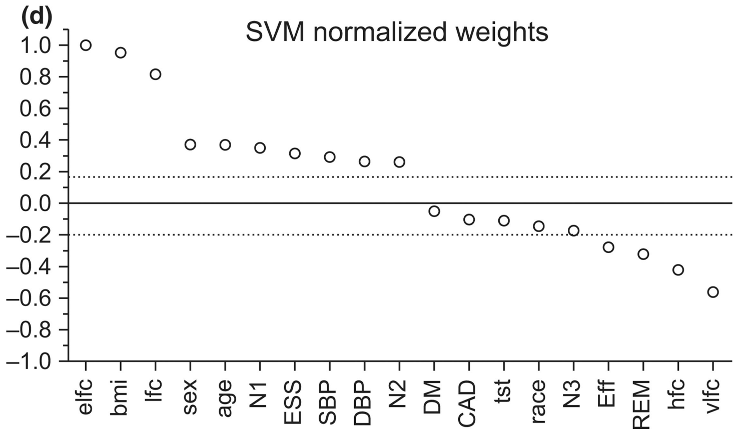


Figure 4.

Performance of the support vector machine (SVM) classifier in predicting apnea–hypopnea index (AHI) class. The prediction power of the SVM algorithm is shown for AHI based on clinical features (a), electroencephalogram (ECG) features (b) or a combination of these and non-respiratory polysomnogram (PSG) features ($n = 19$ features) (c). The parameters gamma, C and epsilon were first searched manually across log-units and then more narrow choices until the shown values were obtained. The parameters used in the data shown were: gamma 0.1, C 0.2, epsilon 0.2. Normalized weights from the SVM are shown in (d), where all values are normalized to the largest weight [elevated low-frequency coupling (e-lfc)]. The absolute value of the weights reflect the strength of the relationship between that feature and the AHI class. The dotted lines mark the range of weighting values obtained when the AHI values were scrambled randomly; feature weights within this range are taken to be non-significant.

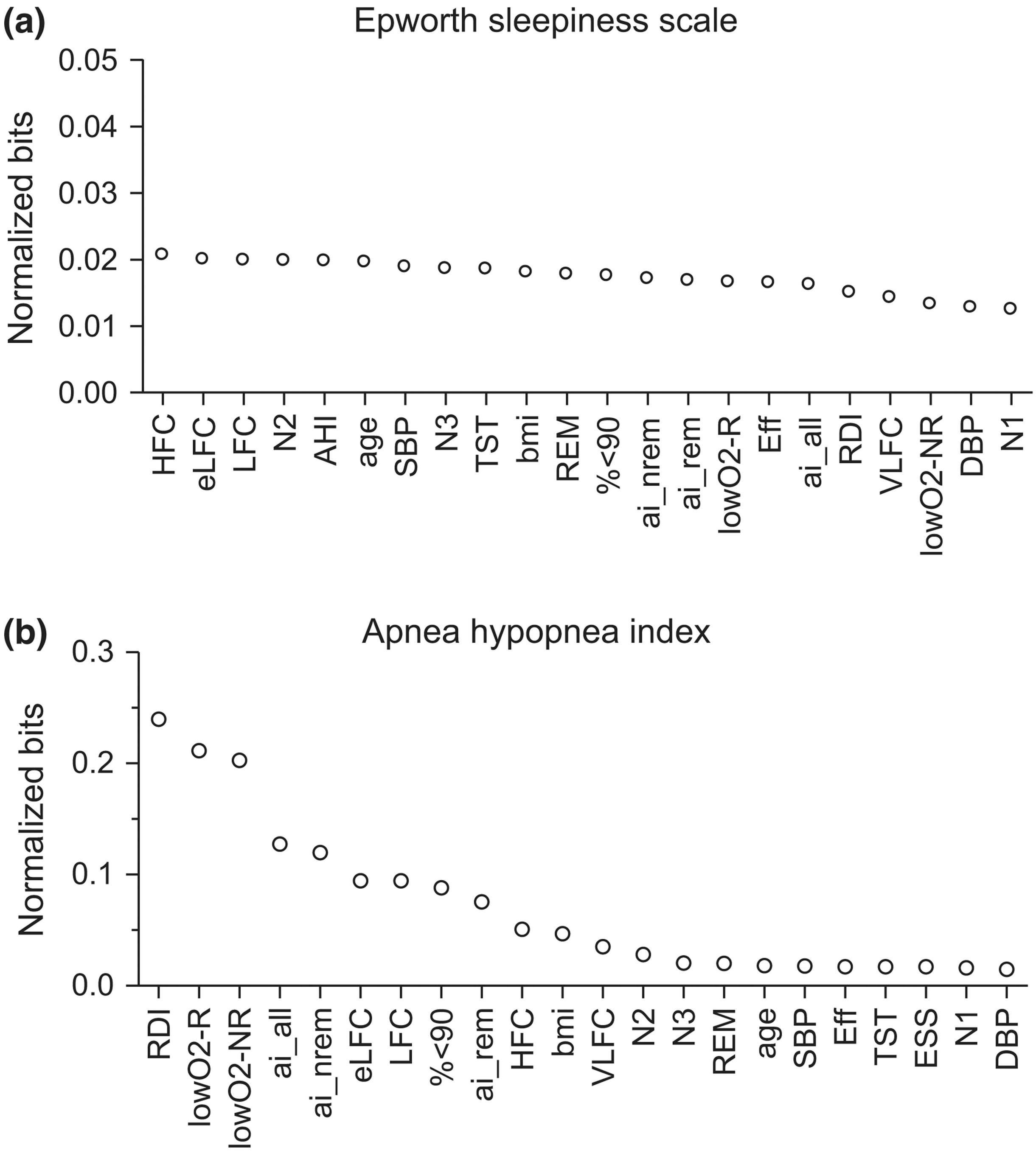


Figure 5.

Mutual information between various features and Epworth Sleepiness Scale (ESS) or apnea–hypopnea index (AHI). The normalized mutual information is shown for various discrete/continuous features compared with ESS (a) and AHI (b). Categorical features were not computed. Note that the vertical axis range is 0–0.05 in (a), while that in (b) is sixfold larger, 0–0.3, emphasizing the striking difference in information between the features and ESS versus AHI. For both plots, the features are stratified from most to least mutual information.

Table 1

Sleep Heart Health Study (SHHS) subject features and abbreviations

HFC: high-frequency coupling	Electrocardiogram (ECG) spectrographic marker of stable,non-fragmented, non-rapid eye movement (NREM) sleep
LFC: low-frequency coupling	ECG spectrographic marker of unstable NREM sleep
VLFC: very low-frequency coupling	ECG spectrographic marker of wakefulness and REM sleep
e-LFC: elevated low-frequency coupling	ECG spectrographic marker of fragmentation, typically from sleep apnea
TST: total sleep time	Total of any stage of sleep in single night of home PSG
N1–3%	NREM stages 1–3, as a percentage of total sleep time
Rapid eye movement (REM)%	REM sleep, as a percentage of total sleep time
Eff: sleep efficiency	TST divided by time in bed from sleep onset until final awakening
ai_all: arousal index	Total number of arousals per hour of sleep
ai_nrem	Number of arousals per hour of NREM sleep
ai_rem	Number of arousals per hour of REM sleep
Time < 90% O_2	% of the TST during which pulse oximetry was < 90%
Low O_2-NR	Lowest pulse oximetry recording during NREM sleep
Low O_2-R	Lowest pulse oximetry recording during REM sleep
RDI	Respiratory disturbance index (per hour of sleep)
AHI	Apnea–hypopnea index (per hour of sleep)
BMI	Body mass index (kg^{-1} m^2)
SBP, DBP	Systolic and diastolic blood pressure (mm Hg)
ESS	Epworth Sleepiness Scale (0–24, where >10 is considered abnormal)
DM	Diabetes mellitus (presence or absence)
CAD	Coronary artery disease (presence or absence)
Sex	Male or female
Age	Years
Race	Caucasian; African American; Native American/Alaskan; Asian/Pacific Islander; Hispanic/Mexican American (left to right in Fig. 1)