pii: jc-19-00144 https://dx.doi.org/10.5664/jcsm.8020

JCSM
Journal of Clinical Sleep Medicine

SCIENTIFIC INVESTIGATIONS

Predicting Nondiagnostic Home Sleep Apnea Tests Using Machine Learning

Robert Stretch, MD[1,2]; Armand Ryden, MD[1,2]; Constance H. Fung, MD, MSHS[1,2]; Joanne Martires, MD[2]; Stephen Liu, MD[2]; Vidhya Balasubramanian, MD[2]; Babak Saedi, MD[2]; Dennis Hwang, MD[3]; Jennifer L. Martin, PhD[1,2]; Nicolás Della Penna, BA[4]; Michelle R. Zeidler, MD, MS[1,2]

[1]David Geffen School of Medicine at University of California, Los Angeles, California; [2]VA Greater Los Angeles Healthcare System, Los Angeles, California; [3]Southern California Permanente Medical Group, Los Angeles, California; [4]Laboratory of Computational Physiology at Massachusetts Institute of Technology, Boston, Massachusetts

Study Objectives: Home sleep apnea testing (HSAT) is an efficient and cost-effective method of diagnosing obstructive sleep apnea (OSA). However, nondiagnostic HSAT necessitates additional tests that erode these benefits, delaying diagnoses and increasing costs. Our objective was to optimize this diagnostic pathway by using predictive modeling to identify patients who should be referred directly to polysomnography (PSG) due to their high probability of nondiagnostic HSAT.

Methods: HSAT performed as the initial test for suspected OSA within the Veterans Administration Greater Los Angeles Healthcare System was analyzed retrospectively. Data were extracted from pre-HSAT questionnaires and the medical record. Tests were diagnostic if there was a respiratory event index (REI) ≥ 5 events/h. Tests with REI < 5 events/h or technical inadequacy—two outcomes requiring additional testing with PSG—were considered nondiagnostic. Standard logistic regression models were compared with models trained using machine learning techniques.

Results: Models were trained using 80% of available data and validated on the remaining 20%. Performance was evaluated using partial area under the precision-recall curve (pAUPRC). Machine learning techniques consistently yielded higher pAUPRC than standard logistic regression, which had pAUPRC of 0.574. The random forest model outperformed all other models (pAUPRC 0.862). Preferred calibration of this model yielded the following: sensitivity 0.46, specificity 0.95, positive predictive value 0.81, negative predictive value 0.80.

Conclusions: Compared with standard logistic regression models, machine learning models improve prediction of patients requiring in-laboratory PSG. These models could be implemented into a clinical decision support tool to help clinicians select the optimal test to diagnose OSA.

Keywords: home sleep apnea testing, machine learning, obstructive sleep apnea, predictive model

Citation: Stretch R, Ryden A, Fung CH, Martires J, Liu S, Balasubramanian V, Saedi B, Hwang D, Martin JL, Della Penna N, Zeidler MR. Predicting nondiagnostic home sleep apnea tests using machine learning. *J Clin Sleep Med*. 2019;15(11):1599–1608.

BRIEF SUMMARY

Current Knowledge/Study Rationale: Patients in whom obstructive sleep apnea was diagnosed through a single home sleep apnea test represent the best-case scenario in terms of balancing the health needs of the patient with the economic and financial impact of sleep testing. Unfortunately, home sleep apnea testing alone is insufficient for a substantial proportion (9% to 35%) of patients. This adds largely underrecognized expense to the process including the direct cost of a second sleep test, as well as the indirect costs of a delayed or missed diagnosis and testing burden on the patient.

Study Impact: This study demonstrates how machine learning can help clinicians to better leverage existing sleep diagnostic modalities to improve patient and healthcare system outcomes without increasing costs.

INTRODUCTION

Obstructive sleep apnea (OSA) is a medical and economic challenge, affecting 4% to 37% of the adult population depending on the diagnostic criteria used and population studied.[1–3] Diagnostic testing for OSA most commonly takes one of the following forms: "attended" in-laboratory polysomnography (PSG) or "unattended" home sleep apnea testing (HSAT) using a portable device. Although HSAT is appealing because of its lower cost and greater convenience for patients, these benefits are only realized if the test is definitive for the diagnosis of OSA. The optimal approach to triaging patients to one of these two tests is not yet fully understood.

Use of HSAT as the first-line diagnostic modality in appropriately selected patients with high pretest probability of moderate to severe OSA has gained widespread acceptance among sleep medicine physicians and accrediting bodies.[4] This has dramatically reduced the need for PSGs. Adoption has been predominantly driven by third-party payor requirements and the lower cost of HSAT. Although the widespread availability of HSAT devices has resulted in greater diagnostic throughput and lower overall costs, 15% to 30%[5,6] of patients cannot be definitively classified as having OSA or not following HSAT. This suggests a suboptimal patient selection process.

Nondiagnostic HSAT occurs when the recording is technically inadequate (ie, due to signal loss) or appears normal with a respiratory event index (REI) < 5 events/h in a patient with suspected OSA. Because of a false-negative rate that may be as high as 17%, the American Academy of Sleep Medicine (AASM) recommends PSG after nondiagnostic HSAT results.[4,7] These guidelines recommend *against* repeat HSAT due to the high likelihood that a second test will also be

nondiagnostic, and increased risk of patients dropping out of the diagnostic process prior to reaching a definitive conclusion.[4,8] For these reasons, indiscriminate use of HSAT carries a risk of harm in the form of delayed diagnoses, missed diagnoses, additional financial burden to the patient and health care system, and misallocation of limited diagnostic resources.

Predictive models are needed to help clinicians determine which patients are unlikely to benefit and could be harmed by attempts to obtain a diagnosis of OSA using HSAT. Although there are two different etiologies of nondiagnostic home sleep apnea tests (HSATs), both necessitate the same management, which is referral for PSG. Training a predictive model to distinguish between these two etiologies is therefore not necessary for the purpose of optimizing the diagnostic pathway. Furthermore, attempts to do so may detract from the predictive accuracy of the model as a consequence of transitioning from binary classification to multinomial classification, which tends to increase model complexity and classification errors. We therefore focused our efforts on developing a binary classifier capable of predicting which patients are likely to have nondiagnostic studies, irrespective of etiology.

Prior studies in a Veterans Administration (VA) population have demonstrated that patients with nondiagnostic HSAT are typically younger, have smaller collar size and body mass index (BMI), are less likely to have hypertension and more likely to carry a diagnosis of posttraumatic stress disorder.[9] Relative to patients with nondiagnostic results due to low REI, those with technically inadequate tests tend to be older with larger collar size and BMI, as well as more likely to have OSA on subsequent PSG (a risk that increases with age across both subtypes of nondiagnostic HSATs).[10] Higher Insomnia Severity Index (ISI) scores correlate with increased risk of a nondiagnostic result on repeat HSAT after an initial nondiagnostic test.[11]

Machine learning offers a promising approach for prediction of nondiagnostic HSATs. Compared to traditional statistical models such as logistic regression, machine learning algorithms offer greater predictive power—albeit often by incurring a decrement in the ability to draw inferences about the relationships between different variables. The aim of this retrospective analysis was to compare machine learning techniques to standard statistical methods in development of a predictive model—based on pre-HSAT questionnaires, patient demographics and comorbidities extracted from the electronic health record—capable of determining which patients will have nondiagnostic HSAT of any type.

METHODS

Data Source and Analytical Packages

Data from patients undergoing HSAT with a Philips Respironics Stardust II (Murrysville, Pennsylvania, United States) device for initial evaluation of suspected OSA at the VA Greater Los Angeles Healthcare System (VA-GLAHS) between October 29, 2013 and July 31, 2014 were analyzed. Pre-HSAT surveys were administered including the following instruments: Berlin questionnaire,[12] STOP-BANG,[13] ISI,[14,15] Epworth Sleepiness Scale,[16] Brief Restless Legs Questionnaire[17] and Patient Health Questionnaire-9 (PHQ-9).[18] Additional patient demographics and health data were extracted from the VistA Computerized Patient Record System (**Table 1**). Patients were excluded if they already carried a sleep disorder diagnosis or had previously undergone any form of diagnostic sleep test. Patients referred for a sleep study at VA-GLAHS are screened by sleep clinicians for contraindications to HSAT in accordance with AASM guidelines, excluding patients with such contraindications from analysis.

All data preprocessing, modeling, and analysis was performed using RStudio (version 1.1.423, RStudio, Inc., Boston, Massachusetts, United States) and the *caret*, *glmnet*, *kernlab*, *randomForest*, *xgboost*, *nnet*, *RANN*, d*iffuStats* and *PRROC* packages.[19–28]

Data Preparation

Each HSAT result was labeled as either diagnostic (obstructive or central sleep apnea diagnosed with REI ≥ 5 events/h) or nondiagnostic (normal or technically inadequate). Hypopneas were defined as ≥ 30% decrease in airflow for ≥ 10 seconds with an associated ≥ 3% oxygen desaturation or an arousal. Clinicians at VA-GLAHS generally adhere to the definition of technical adequacy proposed by Kapur et al[4]: "a minimum of 4 hours of technically adequate oximetry and flow data, obtained during a recording attempt that encompasses the habitual sleep period." If clinically appropriate, physicians could use their professional judgement to deviate from these guidelines. For example, if HSAT with only 3 hours of recording time showed clear evidence of severe OSA, then the study would be deemed diagnostic. This is in line with the aim of our study to facilitate increased diagnostic throughput in clinical practice. The dataset was split into training and testing sets using 4:1 random sampling within each outcome class such that the overall class distribution was maintained. Bivariate analysis of patient characteristics stratified by HSAT outcome was conducted using χ^2 tests (or Fisher exact where appropriate) and Kruskal-Wallis rank-sum tests for categorical and continuous variables, respectively.

Missing data were handled using mixed methods. Data restoration was the preferred approach where possible. For example, the Berlin questionnaire is a screening tool that uses 10 questions across 3 categories to classify patients in a binary manner as "high risk" or "low risk." Depending on which question responses are missing and from which categories those questions came, absent responses are in many cases noncontributory insofar as they would not alter a patient's risk classification when using the questionnaire's validated scoring method. Indeed, some patients are classified as "high risk" by this tool on the basis of as few as 2 of the 10 questions. Using a *minimum viable data* approach, we were able to reconstruct the Berlin classification for most patients and reduce the percent of missing values for this field from 39.5 to 12.4%. Remaining missing values were then handled through imputation using a *k*-nearest neighbor (KNN) algorithm. This enabled the use of model types that lack integrated handling of missing data without necessitating listwise deletion.

Table 1—Candidate input variables for the predictive model.

Manually Sourced		Automatically Sourced		
Patient Responses	**Measurements**	**Comorbidities**	**Demographics**	**Measurements**
Abnormal movements during sleep	Abdominal size	Atrial fibrillation	Age [a]	Body mass index [a]
Berlin questionnaire [a,b]	Collar size [a]	CAD or CHF	Ethnicity [a]	Height [a]
Brief restless legs questionnaire	Hip size	COPD	Sex [a]	Weight [a]
Disturbing dreams or nightmares		Diabetes mellitus	Race [a]	
Epworth Sleepiness Scale [a,b]		Hypertension [a]		
Insomnia Severity Index [a]		CVA		
PHQ-9 [a,b] STOP-BANG [a,b]		PTSD		

Candidate inputs evaluated during model training included manually and automatically sourced variables. The latter refers to those extractable from the electronic health record prior to the patient undergoing their sleep test. [a] Variable included as input to the final model. [b] Subset of questionnaire items included as inputs to the final model. CAD = coronary artery disease, CHF = congestive heart failure, COPD = chronic obstructive pulmonary disease, PHQ-9 = Patient Health Questionnaire-9, PTSD = posttraumatic stress disorder.

Modeling Training and Validation

Candidate input variables were limited to those whose value could be known prior to performing HSAT (**Table 1**). We evaluated several different predictive model types: standard logistic regression, regularized logistic regression, KNNs, random forests, support vector machines, gradient boosted decision trees (GBDTs) and artificial neural networks. For the regularized logistic regression models, we employed two common methods of penalizing large regression coefficients to reduce overfitting: *least absolute shrinkage and selection operator* (LASSO) and *ridge* regression, also known as *L1* and *L2* regularization respectively.[29,30] For models requiring hyperparameter tuning, a grid search approach was employed.

For modeling purposes, a nondiagnostic test was considered a *positive outcome* (ie, correctly predicting that a test would be nondiagnostic was considered a *true positive*). Maximizing both sensitivity (*recall*) and positive predictive value (PPV) (*precision*) were the highest priority. We therefore configured the model training algorithm to optimize the area under the precision-recall curve, which is a metric that balances these two characteristics. More specifically, models optimized the *partial* AUPRC (pAUPRC) where $0 \leq \text{recall} \leq 0.5$ to improve early information retrieval.[31,32]

For model validation, we first performed *k*-fold repeated cross-validation using 10 folds of the training data. Variables were centered, scaled, and missing values imputed within each fold of cross-validation to avoid contamination of training folds, which could increase overfitting. An additional outer validation step was then performed on the 20% holdout testing set. pAUPRC was the primary metric for evaluation and comparison of models.

RESULTS

A total of 613 patients underwent HSAT during the study period. Patient characteristics stratified by HSAT outcome are shown in **Table 2**. Bivariate analyses revealed that patients with nondiagnostic results had significantly lower values for age, weight, BMI, and neck circumference. This difference was even more profound for the subgroup of patients with nondiagnostic results due to normal REI compared with those due to technical inadequacy (**Table S1** in the supplemental material). A greater proportion of patients with nondiagnostic results as compared to those with diagnostic tests self-identified as Hispanic, had higher scores on the PHQ-9 and ISI questionnaires, and lower scores on the STOP-BANG questionnaire. Presence of hypertension, atrial fibrillation, coronary artery disease and congestive heart failure were significantly more common. There was no observed relationship between HSAT outcome and presence of diabetes or total score on ESS.

Data from 491 patients (80.1%) were used for training. The remaining data were set aside for evaluation of the predictive models' performance on data not previously seen. Pre-HSAT sleep-related questionnaires (single item and composite scores) combined with additional patient demographic and health data yielded 62 discrete variables. Of these, 35 were used as inputs in the random forest model, which was the best-performing model overall (**Table 1**). The most important predictors were as follows (in descending order): age, weight, BMI, Berlin questionnaire item on apneas, collar size, height, score on ISI, PHQ-9 question on work impairment, STOP-BANG question on snoring, and ISI question on difficulty falling asleep. Nonlinear models produced using machine learning techniques are often not as readily interpretable as traditional linear regression analyses. In most machine learning models, there are no regression coefficients to indicate the magnitude and directionality of change in the outcome for a given change in an input. We can, however, gain a basic understanding of the relative importance of these inputs to the model in relation to one another (**Figure 1**).

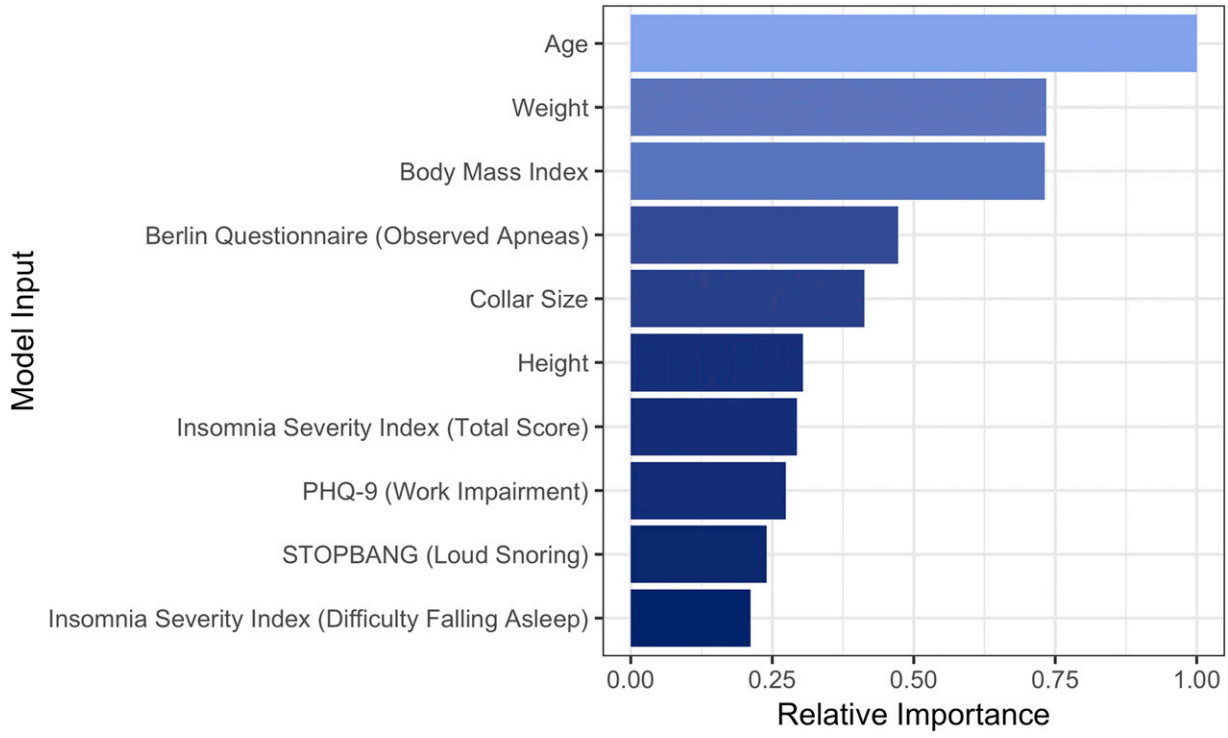
Receiver operating characteristic (ROC) curves plot sensitivity (proportion of disease-positive patients who test positive) against false-positive rate (proportion of disease-negative patients who test positive). Both sensitivity and false-positive rate are prevalence-independent measures. When prevalence differs substantially from 50%, ROC curves can create dramatic visual and quantitative distortion of the test's actual performance.

Table 2—Patient characteristics stratified by outcome of home sleep apnea testing.

Patient Characteristic	Missing (%)	Diagnostic (n = 427)	Nondiagnostic (n = 186)	*P*
Age, years		58.1 [44.2, 66.0]	48.6 [33.3, 63.1]	< .001
Male, %		413 (96.7)	174 (93.5)	.115
Height, inches		70.0 [67.0, 72.0]	70.0 [67.8, 71.5]	.485
Weight, lb		222.0 [196.0, 250.0]	199.0 [180.0, 230.0]	< .001
Body mass index > 35 kg/m^2, %		135 (31.6)	30 (16.1)	< .001
Neck circumference > 17" (male) or 16" (female), %		278 (65.1)	86 (46.2)	< .001
Hispanic, %	12.6	146 (39.6)	94 (56.3)	< .001
Race, %	18.1			< .001
American Indian or Alaska Native		5 (1.4)	5 (3.2)	
Asian		38 (11.0)	35 (22.3)	
Black or African American		68 (19.7)	23 (14.6)	
Native Hawaiian or other Pacific Islander		74 (21.4)	45 (28.7)	
White		160 (46.4)	49 (31.2)	
Comorbidities, %				
Atrial fibrillation		24 (5.6)	3 (1.6)	.045
Cerebrovascular accident		8 (1.9)	2 (1.1)	.731
Chronic obstructive pulmonary disease		20 (4.7)	6 (3.2)	.545
Coronary artery disease or congestive heart failure		49 (11.5)	11 (5.9)	.047
Diabetes mellitus		108 (25.3)	33 (17.7)	.053
Hypertension	0.8	272 (64.0)	83 (45.4)	< .001
Posttraumatic stress disorder		150 (35.1)	74 (39.8)	.313
Epworth Sleepiness Scale	2.8	11.0 [7.0, 16.0]	11.0 [7.0, 15.0]	.996
Epworth Sleepiness Scale, Binned, %	2.8			.462
0–10: Normal daytime sleepiness		197 (47.5)	80 (44.2)	
11–24: Excessive daytime sleepiness		218 (52.5)	101 (55.8)	
Patient Health Questionnaire-9 (PHQ-9) score	10.3	11.0 [6.0, 17.0]	13.0 [7.0, 19.0]	.003
Positive Brief Restless Legs Screen, %	14.4	144 (39.9)	59 (36.0)	.449
STOP-BANG	3.6	5.0 [4.0, 6.0]	4.0 [3.0, 5.0]	< .001
STOP-BANG, Binned, %	3.6			< .001
0–2: Low risk		8 (1.9)	24 (13.9)	
3–4: Intermediate risk		129 (30.9)	78 (45.1)	
5–8: High risk		281 (67.2)	71 (41.0)	
Insomnia Severity Index	8.5	17.0 [12.0, 20.2]	19.0 [15.0, 23.0]	< .001
Insomnia Severity Index, Binned, %	8.5			.003
0–7: No significant insomnia		39 (9.9)	14 (8.3)	
8–14: Subthreshold insomnia		110 (28.1)	26 (15.4)	
15–21: Moderate insomnia		160 (40.8)	76 (45.0)	
22–28: Severe insomnia		83 (21.2)	53 (31.4)	
Home sleep apnea test outcome, %				–
Obstructive sleep apnea		416 (97.4)	0 (0.0)	
Central sleep apnea		11 (2.6)	0 (0.0)	
Normal		0 (0.0)	78 (41.9)	
Technically inadequate		0 (0.0)	108 (58.1)	

Continuous variables are presented as median and interquartile range and statistical testing performed using the Kruskal-Wallis rank-sum test. Categorical variables are presented as n (%) and statistical testing performed using the χ^2 test (or Fisher exact test if the expected cell count was ≤ 5). The percentage of missing values is zero unless otherwise shown.

Figure 1—Relative importance of inputs to the random forest model.



Many machine learning models have their own idiosyncratic method of determining the relative importance of each input. Random forests use *permutation importance*. The 10 most important inputs to the random forest model are shown. Importance values are relative to the most important predictor, in this case patient age, which is assigned a value of 1.0. PHQ-9 = patient health questionnaire.

In contrast, precision-recall curves plot the PPV (*precision*) of the test against sensitivity (*recall*). Because PPV is dependent on disease prevalence, these curves provide better visual and quantitative representations of actual test performance in a population for which disease prevalence is known. Both ROC and precision-recall curves for the most pertinent models are shown in **Figure 2**.

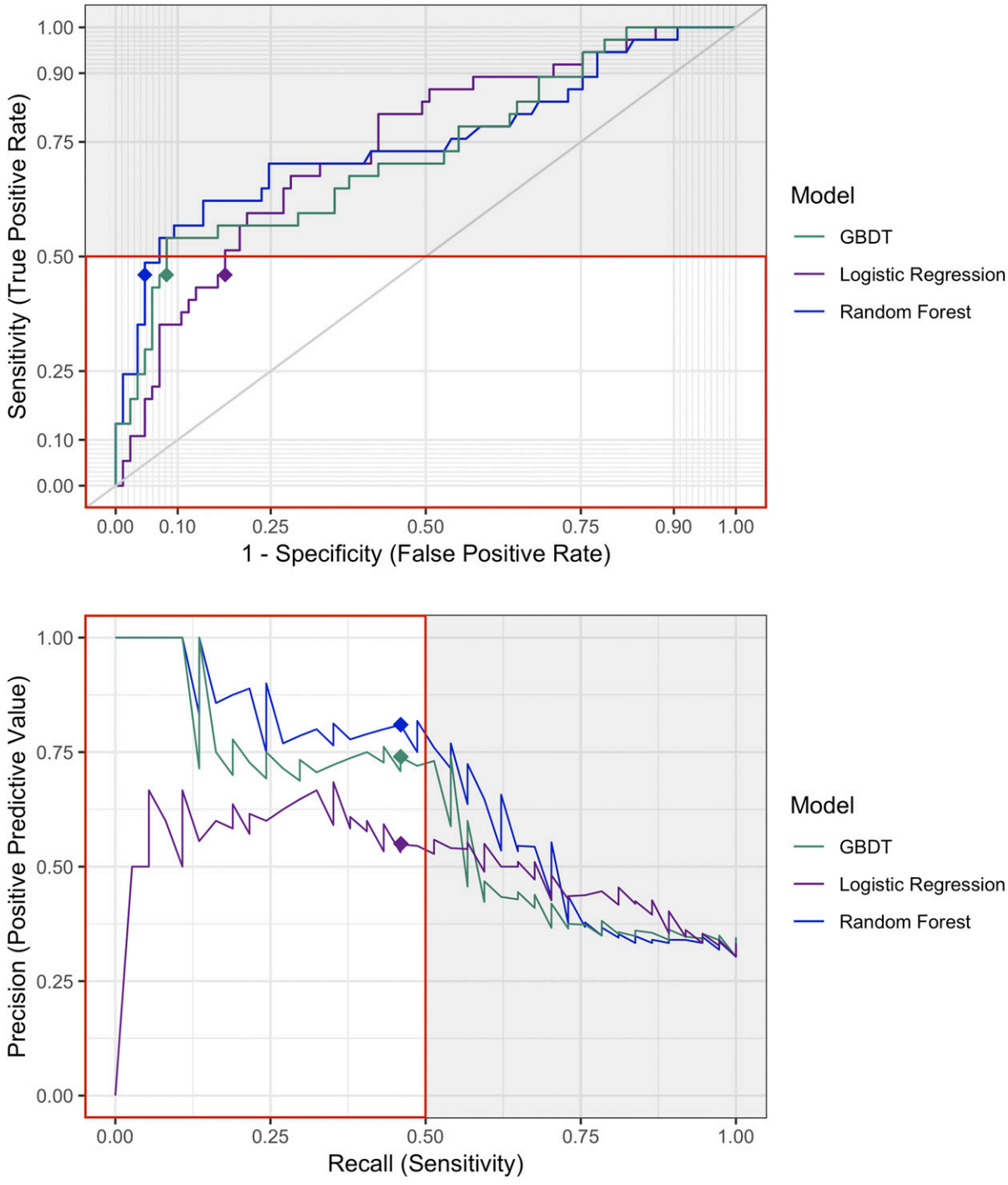
Performance metrics for all models are shown in **Table 3**. Standard logistic regression did not perform as well as other models (pAUPRC 0.574; PPV 0.56 when calibrated to identify 50% of nondiagnostic tests). The penalized regression and artificial neural network models had marginally better performance, as did the support vector machine and KNN models. The GBDT and random forest models both demonstrated substantially better discrimination. The random forest performed best (pAUPRC 0.908; PPV 0.76 when calibrated to identify 50% of nondiagnostic tests). Our preferred calibration for this model yielded sensitivity 0.46, specificity 0.95, PPV 0.81 and negative predictive value 0.80. We can evaluate the effect of implementing this model to guide testing of every 1,000 patients currently being referred for HSAT at the VA-GLAHS (assuming all patients with nondiagnostic HSAT subsequently undergo PSG per AASM guidelines). Without the use of our model, 1,000 patients undergo HSAT; 303 of these patients also undergo PSG due to nondiagnostic HSAT; no patients are referred directly to PSG; total tests performed are HSAT in 1,000 and 303 PSGs. In contrast, when using our model: 828 patients undergo HSAT; 164 of these patients also undergo PSG due to nondiagnostic HSAT; 172 patients are referred directly to PSG; total tests performed are HSAT in 828 and 336 PSGs. In summary, the effects of the model are as follows:

- 139 fewer patients (45.9% decrease) must undergo both HSAT and PSG before obtaining a diagnosis. These patients would instead be diagnosed on the first test.
- 172 fewer home sleep apnea tests (17.2% decrease) at the expense of 33 additional PSGs (10.9% increase). This represents a net cost savings based on published literature demonstrating a twofold to fourfold greater cost of PSG relative to HSAT.[6,33,34]
- 10.5% absolute increase (80.2% up from 69.7%) in the diagnostic yield of HSATs.
- 13.9% absolute increase (83.6% up from 69.7%) in the diagnostic yield of first-time sleep tests (irrespective of whether the first test is by HSAT or a PSG).

DISCUSSION

These findings demonstrate the viability of using machine learning models to predict which patients are likely to require PSG after initial HSAT due to a nondiagnostic result, irrespective of etiology. In existing referral pathways for the diagnosis of OSA, assessment of patient suitability for HSAT depends primarily on the physician's clinical impression of the pretest probability for OSA because no validated clinical decision rules currently exist to serve this purpose.

Figure 2—Receiver operating characteristic (top) and precision-recall (bottom) curves.



Curves for the standard logistic regression, gradient boosted decision tree (GBDT), and random forest models based on holdout test set predictions. A "positive" case or prediction refers to a nondiagnostic home sleep apnea test outcome. Regions with a red border represent the portion of each curve targeted for optimization using the partial area under the precision-recall curve metric. The blue diamond indicates the preferred cutoff point for the random forest model. The green and purple diamonds indicate points with identical sensitivity on the curves representing the other models.

Limited evidence suggests patients referred for HSAT by sleep specialists are somewhat less likely to have a nondiagnostic test than those referred by providers without specialty training (18.7% versus 25.6%, $P < .001$)[5]; however, the nondiagnostic rate remains high even in this subgroup. It is noteworthy that the substantial reduction in nondiagnostic tests seen with our model was achieved in a population that has already been screened by sleep clinicians for any contraindications to HSAT. This suggests there would be an even more dramatic benefit to using similar models in unscreened populations, such as at institutions where clinicians without training in sleep medicine routinely order HSATs (often due to a shortage of sleep clinicians).

Table 3—Model performance on the validation dataset withheld during training.

Model	pAUPRC	Sensitivity	Specificity	PPV	NPV	*F1*-Score
Standard logistic regression	0.574	0.30	0.93	0.65	0.75	0.41
	" "	0.51	0.82	0.56	0.80	0.54
Artificial neural network	0.671	0.30	0.94	0.69	0.75	0.42
	" "	0.49	0.86	0.60	0.79	0.54
Ridge regression	0.673	0.30	0.93	0.65	0.75	0.41
	" "	0.51	0.88	0.66	0.81	0.58
LASSO regression	0.674	0.30	0.92	0.61	0.75	0.40
	" "	0.51	0.86	0.61	0.80	0.56
Support vector machine	0.732	0.30	0.94	0.69	0.75	0.42
	" "	0.51	0.87	0.63	0.80	0.57
k-nearest neighbor	0.766	0.27	0.95	0.71	0.75	0.39
	" "	0.51	0.85	0.59	0.80	0.55
Gradient boosted decision tree	0.801	0.30	0.95	0.73	0.76	0.42
	" "	0.51	0.92	0.73	0.81	0.60
Random forest	0.862	0.30	0.96	0.79	0.76	0.43
	" "	0.51	0.93	0.76	0.81	0.61

Two illustrative cutpoints are shown for each model that approximate sensitivities of 0.30 and 0.50. A "positive" case or prediction refers to a nondiagnostic outcome on home sleep apnea testing. pAUPRC was calculated for the region where $0 \leq \text{recall} \leq 0.5$. LASSO = least absolute shrinkage and selection operator, NPV = negative predictive value, pAUPRC = partial area under the precision-recall curve, PPV = positive predictive value.

The probability of HSAT alone being adequate to diagnose OSA is a function of both the pretest probability of OSA *and* the pretest probability of a nondiagnostic result from any cause. Overreliance on the pretest probability of OSA in the setting of a substantial rate of nondiagnostic tests leads to an erosion of the benefits of HSAT. Instead, we advocate for the use of predictive modeling to facilitate a referral framework that decreases wastage of diagnostic resources without limiting access to testing.

We opted not to delineate between two types of nondiagnostic HSAT (ie, technically inadequate or normal with REI < 5 events/h) because multinomial classification would increase model complexity and the quantity of training data required to achieve the same accuracy. This performance decrement to the model could not be justified because—from the standpoint of clinicians and patients—the relevant endpoint is whether or not HSAT prevents the need for PSG.

Although there is debate regarding the significance of untreated mild OSA, we used a cutoff REI ≥ 5 events/h to classify HSAT as diagnostic because this is both the AASM recommendation and our institutional practice. Alternative REI cutoffs could certainly be used with our model after performing the requisite validation.

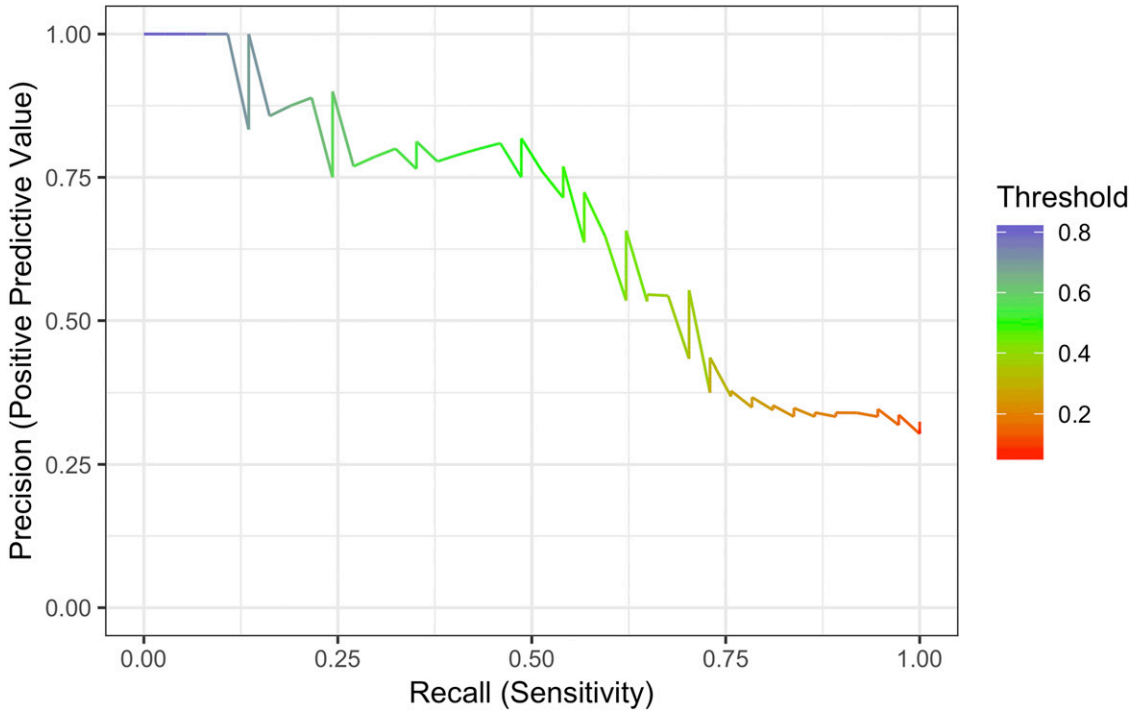
We found that machine learning approaches had superior performance compared with standard statistical techniques. Our results illustrate some of the important limitations of using standard, nonpenalized logistic regression for the type of classification task in this paper. First, the large number of variables in the dataset increases the risk of overfitting, which can lead to a higher level of predictive error when applying the model to a new population. Second, logistic regression is a generalized linear model with only limited flexibility to account for more complex data patterns (eg, through interaction terms and polynomials).

KNN is a versatile nonparametric classification technique. Any given patient's HSAT outcome can be predicted based on the observed outcome for most of the k most similar patients in the training set, where k is a tuned hyperparameter. We found $k = 75$ achieved modest improvements over standard logistic regression. Drawbacks to a KNN approach include computational expense at time of deployment because generating a prediction requires calculating the distances from each new input to every training example in the dataset. This approach also necessitates the training dataset being available on the deployment system.

Decision trees (eg, GBDTs and random forests) are ensemble methods capable of learning nonlinear relationships. Gradient boosting involves training sequential classifiers wherein each classifier corrects for the errors of those before them, minimizing bias and improving predictive accuracy.[35] In contrast, random forests maximize predictive accuracy through reductions in variance rather than bias. Both methods demonstrate robustness against outliers, handle mixed feature types well, and are more appealing than KNN models because they incur minimal system overhead at runtime and do not require the training dataset to remain available during deployment. Furthermore, both have consistently been shown to produce high levels of predictive accuracy across a broad range of problems. It is unsurprising these were the two best performing models in our analysis. Similarly, given the relatively small size of the dataset it could be reasonably anticipated that a neural network architecture would not be able to exceed the predictive accuracy of these tree-based methods.

The results of this study suggest that implementation of a predictive model into an electronic clinical decision support tool (CDST) to help providers choose between ordering HSAT or

Figure 3—Alteration of model characteristics by variation of the cutoff threshold.



The effect on recall and precision of varying the threshold at which the random forest model will predict a given patient's home sleep apnea test will be nondiagnostic.

PSG is capable of delivering meaningful improvements in clinical and economic endpoints. On the basis of all HSAT referrals in this study having been reviewed for appropriateness by sleep physicians, we believe this holds true for providers with or without formal sleep medicine training. Although we reported the effects of our final model using a fixed classification probability threshold (with corresponding sensitivity and specificity), this model could also be implemented using a dynamic threshold (**Figure 3**) wherein the cutoff is automatically adjusted in response to fluctuations in supply and demand for HSATs relative to PSGs within the health care system in which the CDST is deployed. For example, if PSG wait time is substantially lower than usual the threshold could be altered such that the CDST refers a smaller fraction of patients to HSAT than it otherwise would (ie, more patients undergo PSG). Sensitivity (the fraction of patients who would have had nondiagnostic HSAT results that are instead referred for PSG by the CDST) and the number of patients in whom a diagnosis was made on their initial test would both increase, albeit at the expense of a small decrement to the PPV.

One limitation of this study is the relatively small and predominantly male sample sourced from a single health care system. This is of particular importance because women often have milder presentations of OSA than men that may increase the likelihood of a nondiagnostic study. Our follow-up study will include a non-VA health care system among the participating centers and thereby enable a more equal representation of women in the sample. A second limitation is the use of a relatively older HSAT device. Although it was initially thought this older device might have contributed to the higher rate of nondiagnostic studies in our sample (30.3%), our group has since collected unpublished data on 404 consecutive HSATs performed using Vyaire Medical NOX-T3 (Mettawa, Illinois, United States) devices that show a similar nondiagnostic rate (31.9%). We suspect this somewhat higher rate relates to the VA-GLAHS serving a predominantly urban population of Veterans, many of whom have comorbid psychiatric disease, and many with physical disabilities and poor social support. Consistent with this, Tovar Torres et al published nondiagnostic rates as high as 35.1% using the NOX-T3 in their similarly urban Veteran population at the John D. Dingell VA Medical Center in Detroit, Michigan.[36] A final additional limitation is that our model's performance relies on test referral patterns remaining approximately unchanged compared with prior model implementation; otherwise, revalidation may be required; however, this is true of most if not all predictive models and clinical decision rules.

A larger multicenter retrospective analysis utilizing newer HSAT devices is underway to confirm these findings and further refine the predictive model. Our future research plans include the development of a CDST incorporating this refined model, followed by a prospective randomized trial across multiple health care systems to validate the CDST.

ABBREVIATIONS

AASM, American Academy of Sleep Medicine
BMI, body mass index

CDST, clinical decision support tool
ESS, Epworth Sleepiness Scale
GBDT, gradient boosted decision tree
HSAT, home sleep apnea testing
HSATs, home sleep apnea tests
ISI, Insomnia Severity Index
KNN, *k*-nearest neighbor
LASSO, least absolute shrinkage and selection operator
OSA, obstructive sleep apnea
pAUPRC, partial area under the precision-recall curve
PHQ-9, Patient Health Questionnaire
PPV, positive predictive value
PSG, polysomnography
REI, respiratory event index
VA-GLAHS, Veterans Administration-Greater Los Angeles Healthcare System

REFERENCES

1. Peppard PE, Young T, Barnet JH, Palta M, Hagen EW, Hla KM. Increased prevalence of sleep-disordered breathing in adults. *Am J Epidemiol*. 2013;177(9):1006–1014.
2. Oldenburg O, Lamp B, Faber L, Teschler H, Horstkotte D, Töpfer V. Sleep-disordered breathing in patients with symptomatic heart failure: a contemporary study of prevalence in and characteristics of 700 patients. *Eur J Heart Fail*. 2007;9(3):251–257.
3. Schober AK, Neurath MF, Harsch IA. Prevalence of sleep apnoea in diabetic patients. *Clin Respir J*. 2011;5(3):165–172.
4. Kapur VK, Auckley DH, Chowdhuri S, et al. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep Medicine clinical practice guideline. *J Clin Sleep Med*. 2017;13(03):479–504.
5. Al-Sharif H, Culpepper D, Stretch R, Aysola R, Zeidler MR. Outcomes of home sleep apnea testing stratified by specialization of the referring physician. *Sleep*. 2019;42(Suppl_1):A186–A187.
6. Portier F, Portmann A, Czernichow P, et al. Evaluation of home versus laboratory polysomnography in the diagnosis of sleep apnea syndrome. *Am J Respir Crit Care Med*. 2000;162(3):814–818.
7. 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.
8. Rosen CL, Auckley D, Benca R, et al. A multisite randomized trial of portable sleep studies and positive airway pressure autotitration versus laboratory-based polysomnography for the diagnosis and treatment of obstructive sleep apnea: the HomePAP study. *Sleep*. 2012;35(6):757–767.
9. Saedi B, Balasubramanian V, Ryden A, et al. Predictors of a diagnostic home sleep test in veterans. *Sleep*. 2014;37(Abstract Suppl):A131.
10. Zeidler MR, Santiago V, Dzierzewski JM, Mitchell MN, Santiago S, Martin JL. Predictors of obstructive sleep apnea on polysomnography after a technically inadequate or normal home sleep test. *J Clin Sleep Med*. 2015;11(11):1313–1318.
11. Thomas A, Jun D, Ryden A, Zeidler MR. Predictors of a subsequent diagnostic home sleep apnea test after an initial technically inadequate study. Presented at SLEEP 2017, the 31st Annual Meeting of the Associated Professional Sleep Societies; June 3–7, 2017; Boston, MA.
12. Netzer NC, Stoohs RA, Netzer CM, Clark K, Strohl KP. Using the Berlin questionnaire to identify patients at risk for the sleep apnea syndrome. *Ann Intern Med*. 1999;131(7):485.
13. Chung F, Yegneswaran B, Liao P, et al. STOP Questionnaire: a tool to screen patients for obstructive sleep apnea. *Anesthesiology*. 2008;108(5):812–821.
14. Morin CM. *Insomnia: Psychological Assessment and Management*. New York, NY: Guilford Publications; 1996.
15. Bastien CH, Vallières A, Morin CM. Validation of the insomnia severity index as an outcome measure for insomnia research. *Sleep Med*. 2001;2(4):297–307.
16. Johns MW. A new method for measuring daytime sleepiness: the Epworth Sleepiness Scale. *Sleep*. 1991;14(6):540–545.
17. Allen RP, Picchietti D, Hening WA, Trenkwalder C, Walters AS, Montplaisi J. Restless legs syndrome: diagnostic criteria, special considerations, and epidemiology. A report from the restless legs syndrome diagnosis and epidemiology workshop at the National Institutes of Health. *Sleep Med*. 2003;4(2):101–119.
18. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary care evaluation of mental disorders. Patient health questionnaire. *JAMA*. 1999;282(18):1737–1744.
19. RStudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc.; 2015.
20. Kuhn M, Wing J, Weston S, et al. Caret: Classification and Regression Training. R package version 6.0-78. https://CRAN.R-project.org/package=caret. Published 2017. Accessed July 8, 2019.
21. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22.
22. Karatzoglou A, Smola A, Hornik K, Zeileis A. Kernlab: an S4 package for kernel methods in R. *J Stat Softw*. 2004;11(9):1–20.
23. Liaw A, Wiener M. Classification and regression by randomForest. *The Newsletter of the R Project 2*. 2002;3:18–22.
24. Chen T, He T, Benesty M, et al. XGBoost: Extreme Gradient Boosting. R package version 0.71.2. https://CRAN.R-project.org/package=xgboost. Published 2018. Accessed July 8, 2019.
25. Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4th ed. New York, NY: Springer; 2002.
26. Arya S, Mount D, Kemp SE, Jefferis G. RANN: Fast Nearest Neighbour Search (Wraps ANN Library) Using L2 Metric. R package version 2.5.1. https://CRAN.R-project.org/package=RANN. Published 2017. Accessed July 8, 2019.
27. Picart-Armada S, Thompson WK, Buil A, Perera-Lluna A. An R package to compute diffusion-based scores on biological networks: diffuStats. *Bioinformatics*. 2017;btx632.
28. Grau J, Grosse I, Keilwagen J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*. 2015;31(15):2595–2597.
29. Acion L, Kelmansky D, van der Laan M, Sahker E, Jones D, Arndt S. Use of a machine learning framework to predict substance use disorder treatment success. *PLoS One*. 2017;12(4):e0175383.
30. Pajouheshnia R, Pestman WR, Teerenstra S, Groenwold RHH. A computational approach to compare regression modelling strategies in prediction research. *BMC Med Res Methodol*. 2016;16(1):107.
31. Narasimhan H, Agarwal S. A structural SVM based approach for optimizing partial AUC. *PMLR*. 2013;28(1):516–524.
32. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One*. 2015;10(3):e0118432.
33. Pietzsch JB, Garner A, Cipriano LE, Linehan JH. An integrated health-economic analysis of diagnostic and therapeutic strategies in the treatment of moderate-to-severe obstructive sleep apnea. *Sleep*. 2011;34:695–709.
34. Corral J, Sánchez-Quiroga MA, Carmona-Bernal C, et al. Conventional polysomnography is not necessary for the management of most patients with suspected obstructive sleep apnea. Noninferiority, randomized controlled trial. *Am J Respir Crit Care Med*. 2017;196(9):1181–1190.
35. Suen YL, Melville P, Mooney RJ. Combining bias and variance reduction techniques for regression trees. In: Proceedings of the 16th European Conference on Machine Learning. 2005: 741–749.
36. Tovar Torres MP, Salloum A, Sankari A, et al. Determinants for inadequate home sleep apnea testing. *Sleep*. 2016;39(Abstract Suppl):A127.

ACKNOWLEDGMENTS

Author contributions: RS and MZ had full access to all data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. RS performed data cleaning, statistical analysis, and interpretation, as well as writing of the manuscript. NDP aided in the statistical analysis and predictive modeling. CF, AR, DH, and JM contributed to the study design and writing of the manuscript. JB, SL, VB, and BS performed data collection and aided in study design.

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication February 20, 2019
Submitted in final revised form July 8, 2019
Accepted for publication July 12, 2019

Address correspondence to: Michelle R. Zeidler, MD, MS, Greater Los Angeles VA Medical Center, 11301 Wilshire Blvd, Building 500, Room 3025, Los Angeles CA 90073; Email: mzeidler@mednet.ucla.edu

DISCLOSURE STATEMENT

Work for this study was performed at VA Greater Los Angeles Healthcare System. All authors reviewed and were in agreement with the content of the manuscript prior to submission. Research reported in this publication was supported by the American Thoracic Society ASPIRE Award to RS and the Beeson Career Development in Aging Research Award Program (supported by NIA K23AG045937, AFAR, The John A. Hartford Foundation and The Atlantic Philanthropies) to CF. The content is solely the responsibility of the authors and does not necessarily represent the official views of the American Thoracic Society or National Institutes of Health. The authors report no conflicts of interest.