

All subjects provided signed, informed consent before enrollment. The University of Pennsylvania Institutional Review Board approved the protocol.

Subject Selection for Questionnaires

The state of Pennsylvania provided the names of subjects to whom we mailed our questionnaires. This was a random sample of holders of commercial drivers licenses in Pennsylvania within 50 miles of our Sleep Center. Of 4286 questionnaires mailed, we received 1486 (33.0%) responses. Of these, 95 responses were unusable because the CDL holder had died or relocated, so we evaluated our screening strategies among the remaining 1391 (31.5%). Of these, 1329 had complete data to permit calculation of likelihood of apnea [E1] that was used to stratify the sample..

Calculation of Multivariable prediction

Our mailed questionnaire asked: During the past month, have you had, or have you been told about, the following symptom: 1) snorting or gasping, 2) loud snoring, and 3) breathing stops, choking or struggling for breath. Respondents rated symptom occurrence as: Never (0); Rarely, less than once/week (1); Once or twice/week (2); Three or four times/week (3); Five to seven times/week (4); Don't know. We computed the mean symptom-frequency score, with a potential range 0 to 4. When all symptoms were present 5 to 7 times/week, the symptom score was 4. This was the simplest strategy we assessed. We combined this symptom score with BMI, age and gender to determine the multivariable prediction [E1], which ranged 0 to 1, with 0 representing no risk and 1 representing maximal risk for OSA.

Construction of the In-Lab Sample

We used a two-tiered stratified sampling technique [E2, E3] to select respondents for oximetry and polysomnography. We sought to recruit 250 from the 500 respondents at the

highest risk for apnea, and 160 from the remaining, lower-risk respondents, to enrich our final sample studies in the laboratory for the presence of OSA. Our selected sample sizes ensured that the stratum-weighted, pooled proportion with least moderate OSA ($AHI \geq 15$ events/hour) could be estimated with a margin-of-error of $\pm 3\%$, assuming rates of 15% in the higher-risk group and 7% in the lower-risk group. We first sorted the 1329 respondents by the multivariable apnea prediction scores, which has a scale between zero and one (see above) [E1], and invited respondents with the top 500 scores, plus an additional 51 before reaching our recruitment goal. When we reached the driver with the 551st largest score (apnea prediction = 0.436), we had recruited 247 drivers (44.8%). From the remaining 778, we recruited 159 (20.4%) respondents in random order. Thus, 406 subjects (247 high-risk and 159 low-risk) completed in-laboratory studies. We had usable oximetry data on 379 of these. The average \pm SD recording time on oximetry was 6 hours and 47 minutes \pm 47 minutes.

Calculation of Pooled Summary Statistics for In-Lab Sample

When describing the in-lab sample pooled over risk groups, summary statistics were computed as weighted averages of stratum-specific values with standard errors and confidence intervals computed using conventional methods for stratified random sampling. Stratum weights for the higher and lower risk group weights were 0.415 ($=551/1329$) and 0.585 ($=778/1329$), respectively. These weights reflected our estimates of the proportions of commercial driver's license holders at high and low risk in our population, where risk was assigned as high or low depending on whether the multivariable prediction was above or below 0.4356, respectively. Other statistical analyses, such as the ones characterizing the associations between predictive strategy and sleep study data, were already conditional on sample characteristics, and so we did not use sample weights in those analyses.

(25)Scoring of Overnight Oximetry

Continuous, transcutaneous oximetry data were recorded during polysomnography using the N-200 oximeter (Nellcor Inc., Pleasanton, CA) by finger probe. Sampling frequency was 3 Hz, and paper speed 15 cm/hour. We received usable data on 379 studies. The average \pm SD of the duration of the recordings was 7 hours and 32 minutes \pm 47 minutes. A single observer counted desaturations without knowledge of polysomnography results. A *desaturation* occurred if the saturation trace dropped by $\geq 3\%$ below the immediately preceding baseline. Desaturation ended when the level rose by $\geq 2\%$ above the nadir. Finger movement or probe dislodgment artifacts were not counted; these were 1) steep, vertical falls in the trace, followed by steep recoveries, or 2) short, uniform vertical marks 2-4% in magnitude, identifying low quality signal. The oximetry desaturation index (ODI) was the number of desaturations of $\geq 3\%$ magnitude divided by test duration in hours. This first observer and a second re-scored a randomly chosen 10% of traces; intra-class correlation coefficients were computed to assess scoring reliability.

(32)(12)Definition of the Two-Stage Strategy

In our strategy to combine the multivariable prediction and oximetry, we separated in-lab subjects into three groups: those with high, intermediate and low multivariable predictions [E4]. The cutpoints (or parameters) that separated the three groups were variables. We called the value of the multivariable apnea prediction score that separated the high group from the intermediate the “upper bound”, while the score separating the intermediate from the lower group was the “lower bound”. Those with scores in the high range were predicted to have OSA, with subsequent review of their sleep study to assess this prediction. Those with scores in the lower range would be predicted to be free of OSA and would not be further studied. We assessed the result of their sleep study to determine whether our prediction of no sleep apnea was correct.

Those with scores in the intermediate range would, in this strategy, undergo oximetry; this group's desaturation indices were compared against a threshold value, a variable called the ODI threshold. Those with desaturation indices exceeding this threshold would be predicted to have OSA and hence would undergo polysomnography, while the rest would be predicted to be free of OSA (see Figure 1). We would assess their sleep study results to determine the correctness of our prediction. We determined the optimal values for these three variables (upper bound, lower bound, desaturation threshold) for predicting severe apnea [E5], and as a secondary objective, a different set of values for any apnea [E5].

Determination of Discriminatory Power

We performed "Area Under the Curve" (AUC) analysis for receiver operating characteristic (ROC) curves [E6, E7], using ROCKIT (Chicago, IL) [E8]. This analysis determined the relative discriminatory power of symptoms, BMI, multivariable prediction, and oximetry (34). ROC curves were also constructed for the two-stage strategy, by plotting sensitivity against 1-specificity (see below). Based on published methods [E9], optimal sensitivity and specificity were identified by taking the values associated with the point on the ROC where the tangent line's slope equaled $[(1-p)/p][FP/FN]$, where p = apnea prevalence, FP = false positive rate, and FN = false negative rate. We believed that missing cases of *severe* apnea was much more costly than a false positive, a reasonable assumption for screening applications with public safety implications, and so we weighted FP/FN at 1:20. For predicting *any* apnea, we weighted this ratio at 1:3, assuming that the cost of missing a case of *any* apnea was less serious. The estimated proportion of *any* apnea in our population was 0.281 and the proportion of *severe* apnea was 0.047. Thus, for $AHI \geq 30/\text{hour}$, the slope = $(1-0.047)/(0.047)(1/20) \sim 1$ and for $AHI \geq 5/\text{hour}$, the slope = $(1-0.281)/(0.281)(1/3) \sim 1$.

Building the ROC Curve for the Two-Stage Screen

For each of our two objectives, we varied upper bound of the multivariable prediction score from 0.2 to 0.9, in 0.1 unit increments. We varied lower bound from 0.1 to the value of upper bound minus 0.1. Thus, when upper bound was 0.9, there were 8 values of lower bound (0.1 to 0.8, in 0.1-unit increments). When upper bound was 0.8, there were 7 values of lower bound, and so on. We varied threshold ODI from 5 to 25 events/hour in increments of 5 events/hour, as described previously [E4]. Using SAS programming (Cary, NC), for each of the $[(8+7+6+5+4+3+2+1) \times 5] = 180$ combinations of these three variables, we computed sensitivity and specificity for predicting severe apnea, and secondarily, at least mild apnea.

We generated a receiver-operating-characteristic curve for the two-stage strategy by plotting sensitivity against 1-specificity. We note that several sensitivity values could be associated with a unique specificity. To address this issue, we rank-ordered specificity, then selected the highest corresponding cutpoint value associated with the non-unique sensitivities associated with that specificity value. We plotted sensitivity against 1-specificity, and computed AUC using SAS. We determined the optimal sensitivity and specificity as above, by selecting the sensitivity and specificity associated with the point on the ROC curve whose tangent line gave unit slope. The parameter combination associated with this value of sensitivity and specificity was the optimum parameter set. We reported this sensitivity and specificity in Table 2, and we computed and reported negative likelihood ratios as $(1 - \text{sensitivity}) / \text{specificity}$.

Calculation of Confidence Intervals for the Single-Stage Strategies

Using the SAS jackboot macro, we performed bootstrap re-sampling [E10] to generate non-parametric 95% confidence intervals around the estimates of AUC, sensitivity, specificity, and negative likelihood ratios shown in Table 2.

For our single-stage strategies, we re-sampled the in-lab data on 406 participants regarding symptoms, BMI, multivariable predictions, and oxyhemoglobin desaturation indices. We computed AUC's as the c-statistic generated from a SAS logistic regression on 1000 bootstrap re-samples. We used the optimal cutpoint for each strategy to compute sensitivity, specificity, and negative likelihood ratios on each of 1000 bootstrap re-samples, creating a bootstrap distribution of values for each of these estimates. We then computed nonparametric 95% lower and upper confidence limits using the 2.5th and 97.5th percentile values of each of these three (sensitivity, specificity and negative likelihood ratio) distributions.

Calculation of Confidence Intervals for the Two-Stage Strategies: AUC

For the two-stage strategy, we constructed confidence intervals around AUC in the following way. First, we re-sampled from the in-lab data (multivariable prediction, oximetry and AHI values), and from each, we applied our 180 sets of upper bound, lower bound and ODI threshold. Doing so provided 180 values of sensitivity and 1-specificity for each re-sample, for each of our two AHI criteria (≥ 5 /hour and ≥ 30 /hour), from which we built 1000 ROC curves. The AUC was computed using the SAS "area" macro for each of these curves. As for the single-stage strategies, we determined nonparametric 95% lower and upper confidence limits using the 2.5th and 97.5th percentile values of AUC.

Calculation of Confidence Intervals for the Two-Stage Strategies: Sensitivity, Specificity and Negative Likelihood

To compute confidence intervals around sensitivity, specificity and negative likelihood for the two-stage strategy, we first selected the optimum cutpoints for ODI threshold, upper bound and lower bound, as described above. We re-sampled values of BMI, multivariable prediction and ODI threshold, and applied these cutpoints to each set of re-sampled data. Doing

so gave us one value of sensitivity, specificity and negative likelihood. Re-sampling 1000 times gave us 1000 such values, and we determined the 2.5th and 95.5th percentile values of this distribution to determine the nonparametric 95% lower and upper confidence limits around the observed values.

REFERENCES

- E1. Maislin G, Pack A, Kribbs N, Smith P, Schwartz A, Kline L, Schwab R, and Dinges D. A survey screen for prediction of apnea. *Sleep* 1995;18:158-166.
- E2. Young T, Palta M, Dempsey J, Skatrud J, Weber S, and Badr S. The occurrence of sleep-disordered breathing among middle-aged adults. *New Engl J Med* 1993;328:1230-1235.
- E3. Gislason T, Almqvist M, Eriksson G, Taube A, and Boman G. Prevalence of sleep apnea syndrome among Swedish men--an epidemiological study. *J Clin Epidemiol* 1988; 41(6):571-576.
- E4. Gurubhagavatula I, Maislin G, and Pack AI. An algorithm to stratify sleep apnea risk in a sleep disorders clinic population. *American Journal of Respiratory & Critical Care Medicine* 2001;164(10):1904-1909
- E5. Sleep-Related Breathing Disorders in Adults: Recommendations for Syndrome Definition and Measurement Techniques in Clinical Research. The Report of an American Academy of Sleep Medicine Task Force. *Sleep* 1999; 22(5):667-689.
- E6. Hanley J, and McNeil B. The meaning and use of area under a receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29-36.
- E7. Harrell F, Lee K, and Mark D. Tutorial in biostatistics: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361-387.
- E8. ROCKIT Software. Department of Radiology, Biological Sciences Division. University of Chicago Hospitals and Clinics. Chicago, Illinois. Available ftp at random.bsd.uchicago.edu 1998

- E9. McNeil B, Keeler E, and Adelstein S. Primer on certain elements of medical decision making. *New Engl J Med* 1975;293(5):211-215.
- E10. Efron B, and Tibshirani R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Stat Sci* 1986;1:54-77.