Anesthesiology 2009; 110:928–39

A Meta-analysis of Clinical Screening Tests for Obstructive Sleep Apnea

Satya Krishna Ramachandran, M.D., F.R.C.A.,* Lydia A. Josephs, M.D.†

The purpose of this meta-analysis is to compare clinical screening tests for obstructive sleep apnea and establish an evidence base for their preoperative use. Diagnostic odds ratios were used as summary measures of accuracy, and false-negative rates were used as measures of missed diagnosis with each screening test in this review. Metaregression revealed that clinical models, logarithmic equations, combined techniques, cephalometry, and morphometry are significant characteristics, whereas body mass index, history of hypertension, and nocturnal choking are significant test elements associated with higher diagnostic accuracy. Test accuracy in repeated validation studies of the same screening test is variable, suggesting an underlying heterogeneity in either the clinical presentation of obstructive sleep apnea or the measured clinical elements of these models. Based on the false-negative rates, it is likely that most of the clinical screening tests will miss a significant proportion of patients with obstructive sleep apnea.

OBSTRUCTIVE sleep apnea (OSA) affects 2–4% of the population[1] in the United States and is now considered a significant risk factor for perioperative morbidity and mortality.[2,3] The risks of OSA in the general population are well known and include hypertension,[4] coronary artery disease,[5,6] stroke,[7] pulmonary hypertension,[8] sudden cardiac death,[9] and deep vein thrombosis,[10] to name a few that directly impact on perioperative outcome. Overnight polysomnography is the standard for diagnosis of OSA, but its value in the management of patients scheduled to undergo surgery is reduced by significant issues with resource availability.[11] Full polysomnography involves an overnight stay in a designated sleep laboratory with multichannel monitoring to measure electro-oculogram, chin and leg electromyography, electro-oculography, chest and abdominal respiratory effort, nasal airflow *via* a thermistor and/or nasal cannula, oxygen saturation, and heart rate monitoring, in addition to several sleep architecture measures. Traditionally, the apnea–hypopnea index (AHI) or the respiratory disturbance index has been used as a measure of the presence of OSA and its severity. Accepted diagnostic thresholds for OSA have varied between AHI values of 5 or more per hour[1,12] and 10 or more per hour.[13]

Current guidelines by the American Society of Anesthesiologists (ASA) recommend preoperative polysomnography when indicated.[12] Although it may indeed be the most cost-effective strategy in diagnosing OSA,[14] urgency of the planned operative procedure is an important limiting factor in pursuing a policy of liberal preoperative polysomnography.[15] It is estimated that 93% of females and 82% of males with OSA are possibly undiagnosed.[16] Further, it will take several years to complete the current requirement for polysomnography in the general population with existing resources.[11] All of these points make a compelling argument in favor of cost-effective prediction models to help anesthesiologists assess risk of OSA preoperatively. It is with this background information that we set out to systematically review alternatives to polysomnography in published literature. Indeed, there have been numerous efforts in the past to devise alternate clinical methods of predicting OSA, primarily by experts in sleep medicine, looking to aid in screening patients for high risk of OSA. These methods are broadly classified as questionnaires and clinical prediction models (algorithms, artificial neural networks, cephalometry, morphometry, and other combined techniques and regression models). There is no consensus in the ASA or the American Academy of Sleep Medicine about the best screening tests, with the exception of portable devices for diagnosis of OSA.[15] Most of the current screening methods have been validated in the sleep laboratory population. It is important to recognize that basic differences exist between the study populations in sleep laboratories and preoperatively. On the one hand, patients are referred to sleep laboratory because of a perceived high risk of OSA, and a questionnaire or clinical screening test administered in the sleep laboratory essentially functions as a second highly specific step to rule in the diagnosis of OSA. Anesthesiologists, on the other hand, need a highly accurate clinical test with high sensitivity to rule out OSA robustly in a lower-risk population, without recourse to confirmatory polysomnography. In addition, screening test results from high-risk populations often report higher sensitivity than is seen when the test is used in a lower-risk population. Identifying the most accurate screening test, with reproducible low false-negative rates, is of critical importance in this context. A previous meta-analysis on screening tests for OSA[17] was published in 2000 and described the analysis of several screening methods, including partial time polysomnography, partial channel polysomnography, oximetry, portable devices, prediction equations, flow–volume loops, global impression, questionnaires, and other clinical, chemical, and radiologic screening tests. Methodologically, it suffers from a major inade-

* Clinical Lecturer, † Medical Student.

Received from the Department of Anesthesiology, University Hospital, Ann Arbor, Michigan. Submitted for publication June 6, 2008. Accepted for publication December 10, 2008. Support for this work was provided solely from institutional and/or departmental sources.

Mark A. Warner, M.D., served as Handling Editor for this article.

Address correspondence to Dr. Ramachandran: Department of Anesthesiology, 1 H427 University Hospital Box 0048, 1500 E Medical Center Drive, Ann Arbor, Michigan 48109-0048. rsatyak@med.umich.edu.

quacy in that several largely heterogeneous studies were pooled without analysis of relative merits and demerits of the individual tests. This lack of discriminatory analysis makes it difficult for an anesthesiologist to make an evidence-based choice of preoperative screening test. Further, there have been several new validation studies on prediction models for OSA in the ensuing years after the last meta-analysis. The purpose of the current systematic review is, therefore, to update the literature and identify the best approach to clinical prediction of OSA, by comparing clinical screening tests for ease and accuracy of prediction of both diagnosis and severity of OSA. We investigated this question using quantitative methods to retrieve and analyze the relevant published literature.

Materials and Methods

The reviewers (S.K.R. and L.A.J.) searched the electronic databases PubMed and Ovid for articles published in English from 1966 to May 2008 using the phrases *sleep apnea*, *obstructive sleep apnea*, *prediction*, *diagnosis*, *screening*, and combinations of these phrases. We then manually searched the associated articles and bibliography of any relevant published article we retrieved for additional pertinent references. We also checked the Cochrane Controlled Trials register and hand searched the journals *Sleep*, *American Journal of Respiratory and Critical Care Medicine*, *Thorax*, *Chest*, *International Journal of Obesity*, *Obesity*, *Obesity Reviews*, *Obesity Surgery*, and *Annals of Internal Medicine* for additional articles.

Inclusion Criteria and Assessment of Study Quality

Studies that measured the diagnostic value of questionnaires, clinical scales, or prediction equations (algorithms or regression equations) compared with standard overnight polysomnography were included. We excluded from review studies that did not provide prevalence (pretest probability) of OSA with raw data in 2 × 2 tables, sensitivity and specificity, or positive and negative likelihood ratios. We also excluded studies where the reference standard was not overnight monitored polysomnography in a hospital or laboratory facility. We therefore excluded studies that used portable monitoring as the standard. Two reviewers (S.K.R. and L.A.J.) independently screened the titles and abstracts of all articles identified by the search strategy and individually determined inclusion of studies for the analysis, guided by previously established methodologic standards for diagnostic test research in OSA.[18] Full copies of all selected articles were retrieved. Disagreements regarding inclusion and exclusion were resolved by discussion, which involved manually rechecking the data extraction from each disputed article. In those situations where there was lack of agreement or clarity even after this discussion, further advice was sought from our institution's specialist on OSA (Ronald D. Chervin, M.D.). The quality of studies accepted for this review was analyzed under the Quality Assessment of Diagnostic Accuracy Studies[19] framework for completeness and accuracy of reporting.

Definitions and Statistical Analysis

A questionnaire was defined as a set of questions with no additional physical measurement involved. A clinical model combined elements of history and physical examination, with or without additional measurements and investigations (radiologic, oximetry, or laboratory). Of these, we arbitrarily chose three elements to define ease of use of a given test and reflect its applicability as a preoperative screening tool: number of variables, use of linear or log scales, and description of the clinical methods. An ease-of-use scale was thus developed, with 0 defining easy and 3 meaning complex methodology. One point was added for the presence of each of the following: four or more test elements or variables; log scale; and need for additional techniques, measurements, or investigations.

The frequency of true-positive, true-negative, false-positive, and false-negative (FN) results were abstracted from all selected studies. True positives were defined as the frequency of patients with OSA with a positive screening test. False positives were the frequency of patients with positive screening test but no OSA. FNs were the frequency of patients with OSA and a negative screening test. True negatives referred to the frequency of patients with negative screening test and no OSA. Where raw 2 × 2 tables were not presented, these data were derived from the relevant results. From each 2 × 2 table, we computed sensitivity, specificity, likelihood ratios, and the diagnostic odds ratio (DOR), which combines data on sensitivity and specificity to give an indication of a test's ability to rule in or rule out a condition. The DOR was chosen as the primary summary measure of test accuracy for comparison, the reasons for which are described in further detail in the Discussion. A DOR of greater than 81 was chosen to identify an excellent test, because this indicated that the specificity and sensitivity were both greater than 0.9.[20] A DOR of 10–80 was termed a good test,[21] 5–10 was arbitrarily termed average, 2–5 was considered poor, and less than 2 was considered to be of no value in prediction. These summary measures of diagnostic accuracy were reported as point estimates with 95% confidence intervals. The FN rate was derived as (1 − sensitivity) and was used as a measure of the rate of missed diagnosis for any given screening test. An ideal test was one that had a DOR greater than 81 and an FN rate of 0%.

In addition, the following unique descriptors of the test were collected: questionnaire or clinical model, year of publication, Quality Assessment of Diagnostic Accu-

racy Studies score, number of patients in the study, age, sex balance, prevalence of OSA, mean body mass index, number of variables, linear scale or log equation, and presence of additional diagnostic modalities (cephalometry, morphometry, or oximetry). Elements of the screening tests were also collected as binary yes/no data, including body mass index, snoring, age, sex, hypertension, witnessed apnea, neck circumference, choking or gasping in sleep, tiredness, and daytime somnolence. We computed statistics for individual studies and combined them using Meta-DiSc (version 1.2; Ramon y Cajal Hospital, Madrid, Spain).[22] We planned to use the Mantel–Haenszel fixed effects model if the studies were homogeneous for the diagnostic performance indices and the DerSimonian–Laird random effect model if they showed heterogeneity. A further within-test subanalysis was performed for screening tests with more than one validated study, to assess whether the reported accuracy of the test in one study was reproducible across all studies. The homogeneity of likelihood ratios and DORs were traditionally assessed using the Cochran Q test based on inverse variance weights, which also has a chi-square distribution with $k - 1$ degrees of freedom.[22] A P value less than 0.05 was considered to indicate the presence of statistically significant heterogeneity between the studies. In addition, the I^2 index was used to quantify any heterogeneity. A value of 0% indicates no heterogeneity, and larger values indicate increasing heterogeneity. Low, moderate, and high I^2 values of 25, 50, and 75% were chosen to quantify heterogeneity as described by Higgins *et al.*[23]

Finally, a random effects metaregression of the previously listed screening test elements was undertaken, by adding these elements as covariates to the regression model. The resulting parameter estimates were back-transformed (antilogarithmic transformation) to relative diagnostic odds ratios (rDORs).[24] An rDOR of 1 indicates that the particular screening test element does not affect the overall DOR of the test. An rDOR of greater than 1 means a particular element bestows a higher DOR on a test, compared with tests without this particular variable. For the purposes of description, rDOR > 2 was arbitrarily chosen to identify significant variables. Diagnostic threshold effect was studied using Littenberg and Moses' fitted model,[25] $D = a + bS$, where D is the natural logarithm of the DOR and S is the natural logarithm of the product of the odds of true-positive test results and the odds of false-positive test results. The statistical significance of the regression coefficient b ($P < 0.05$) was tested to assess whether diagnostic accuracy varies significantly with changes in threshold.

Results

Our initial search strategy (fig. 1) retrieved 6,816 potentially relevant diagnostic studies, which were then screened by title first and then by abstract. These included the unduplicated results of multiple search engines and hand-searched journals as listed previously. After review of the relevant articles and their bibliography, 115 studies were considered potentially appropriate for the study. On a more detailed review of these publications, a further 89 articles were excluded from the final analysis, either because the standard reference test used was not overnight polysomnography or clinical methods were not used in the screening tests. Therefore, a total of 26 articles were accepted for final analysis. Of these, 8 pertained to questionnaires, and the remaining 18 described clinical prediction tests. These included linear scales, algorithms, regression models, morphometry, cephalometry, combined prediction models, and neural networks. The 26 studies included a total 6,794 patients with suspected OSA (median sample size, 123; range, 33–1,409). The study prevalence of OSA ranged from 0.09 to 0.847. The proportion of male subjects ranged from 24% to 100%.

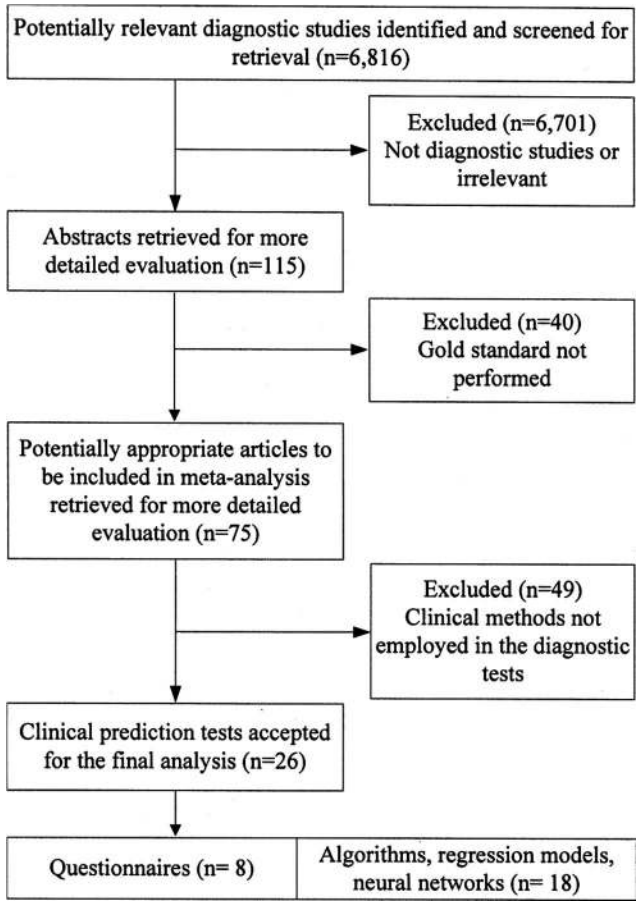

Fig. 1. Flow diagram of the systematic review process.

The study quality of the included articles was variable with Quality Assessment of Diagnostic Accuracy Studies scores ranging from 6 to 13. In addition, there was evidence of verification bias as described by Irwig *et al.*,[18] because several studies were derived from nonrandomly chosen populations and did not describe the cases that

Table 1. Test Characteristics of Questionnaires Predicting the Diagnosis of Obstructive Sleep Apnea

Study	Prevalence of OSA	AHI Threshold	Study n	Sensitivity (95% CI)	Specificity (95% CI)	LR^+ (95% CI)	LR^- (95% CI)
ASA checklist; Chung *et al.*,[38] 2008	0.696	5	177	0.721 (0.633–0.799)	0.382 (0.254–0.523)	1.167 (0.922–1.476)	0.730 (0.470–1.135)
Berlin questionnaire; Ahmadi *et al.*,[39] 2008	0.262	5	130	0.676 (0.495–0.826)	0.490 (0.386–0.594)	1.325 (0.978–1.796)	0.661 (0.390–1.120)
Berlin questionnaire; Ahmadi *et al.*,[39] 2008	0.262	10	130	0.618 (0.436–0.778)	0.427 (0.327–0.532)	1.078 (0.786–1.479)	0.895 (0.551–1.456)
Berlin questionnaire; Sharma *et al.*,[40] 2006	0.596	5	104	0.855 (0.742–0.931)	0.952 (0.838–0.994)	17.952 (4.624–69.691)	0.152 (0.083–0.280)
Berlin questionnaire; Sharma *et al.*,[40] 2006	0.596	10	104	0.855 (0.742–0.931)	0.857 (0.715–0.946)	5.984 (2.833–12.641)	0.169 (0.091–0.314)
Berlin questionnaire; Chung *et al.*,[38] 2008	0.696	5	177	0.689 (0.598–0.769)	0.545 (0.406–0.680)	1.515 (1.108–2.072)	0.571 (0.399–0.816)
Epworth Sleepiness Scale; Osman *et al.*,[41] 1999	0.457	5	46	0.286 (0.113–0.522)	0.520 (0.313–0.722)	0.595 (0.270–1.311)	1.374 (0.864–2.184)
Sleep questionnaire; Haraldsson *et al.*,[42] 1992	0.429	10	42	0.778 (0.524–0.936)	0.792 (0.578–0.929)	3.733 (1.647–8.460)	0.281 (0.115–0.682)
SDQ females; Weatherwax *et al.*,[43] 2003	0.552	5	55	0.800 (0.593–0.932)	0.667 (0.472–0.827)	2.400 (1.395–4.130)	0.300 (0.132–0.684)
SDQ males; Weatherwax *et al.*,[43] 2003	0.552	5	70	0.750 (0.597–0.868)	0.654 (0.443–0.828)	2.167 (1.244–3.775)	0.382 (0.213–0.685)
Snoring questionnaire; Bliwise *et al.*,[13] 1991	0.423	10	1,409	0.304 (0.267–0.342)	0.988 (0.977–0.994)	24.690 (13.178–46.259)	0.705 (0.668–0.744)
STOP questionnaire; Chung *et al.*,[31] 2008	0.696	5	177	0.656 (0.564–0.739)	0.600 (0.459–0.730)	1.639 (1.157–2.322)	0.574 (0.414–0.795)
Symptoms; Gurubhagavatula *et al.*,[44] 2004	0.28	5	406	0.518 (0.422–0.612)	0.685 (0.628–0.738)	1.643 (1.286–2.099)	0.704 (0.574–0.865)
Pooled estimates	0.28–0.696			0.520 (0.493–0.546)	0.800 (0.779–0.819)	2.468 (2.210–2.757)	0.642 (0.608–0.678)

AHI threshold = diagnostic threshold of apnea–hypopnea index used in each study; ASA = American Society of Anesthesiologists; CI = confidence interval; LR^+ = positive likelihood ratio; LR^- = negative likelihood ratio; OSA = obstructive sleep apnea; SDQ = Sleep Disorders Questionnaire; STOP = acronym from Chung *et al.*[31]

were not included in sufficient detail. Tables 1–4 describe the sensitivity, specificity, and likelihood ratios of the various screening tests for prediction of OSA. Random effects models were used because of the significant heterogeneity between studies, as measured by both the Cochran Q test ($P < 0.05$) and the I^2 index (74.4–90.6%).

An additional subgroup analysis was undertaken of the screening tests with more than one validation study. Three tests were identified for this purpose, namely the Berlin questionnaire, the Maislin multivariable apnea index (algorithm), and the Kushida index (morphometry). There was a high degree of heterogeneity within each

Table 2. Test Characteristics of Questionnaires Predicting the Presence of Severe Obstructive Sleep Apnea

Study	Prevalence of OSA	AHI Threshold	Study n	Sensitivity (95% CI)	Specificity (95% CI)	LR^+ (95% CI)	LR^- (95% CI)
ASA checklist; Chung *et al.*,[38] 2008	0.696	30	177	0.877 (0.805–0.930)	0.364 (0.238–0.504)	1.378 (1.117–1.701)	0.338 (0.188–0.609)
Berlin questionnaire; Sharma *et al.*,[40] 2006	0.596	30	104	0.919 (0.822–0.973)	0.667 (0.505–0.804)	2.758 (1.787–4.257)	0.121 (0.051–0.288)
Berlin questionnaire; Chung *et al.*,[38] 2008	0.696	30	177	0.870 (0.797–0.924)	0.463 (0.326–0.604)	1.620 (1.253–2.094)	0.28 (0.164–0.482)
Sleep questionnaire; Haraldsson *et al.*,[42] 1992	0.429	25	43	0.895 (0.669–0.987)	0.750 (0.533–0.902)	3.579 (1.760–7.279)	0.140 (0.037–0.531)
SDQ females; Douglass *et al.*,[45] 1994	0.090	25	187	0.882 (0.636–0.985)	0.812 (0.745–0.868)	4.688 (3.280–6.700)	0.145 (0.039–0.534)
SDQ males; Douglass *et al.*,[45] 1994	0.424	25	332	0.851 (0.781–0.905)	0.759 (0.692–0.818)	3.534 (2.722–4.588)	0.196 (0.131–0.293)
STOP questionnaire; Chung *et al.*,[31] 2008	0.696	30	177	0.795 (0.713–0.863)	0.491 (0.354–0.629)	1.562 (1.187–2.056)	0.417 (0.269–0.649)
Pooled estimates	0.09–0.696			0.858 (0.828–0.885)	0.679 (0.639–0.71)	2.177 (1.941–2.441)	0.231 (0.183–0.292)

AHI threshold = diagnostic threshold of apnea–hypopnea index used in each study; ASA = American Society of Anesthesiologists; CI = confidence interval; LR^+ = positive likelihood ratio; LR^- = negative likelihood ratio; OSA = obstructive sleep apnea; SDQ = Sleep Disorders Questionnaire; STOP = acronym from Chung *et al.*[31]

Table 3. Test Characteristics of Clinical Models Predicting the Diagnosis of Obstructive Sleep Apnea

Clinical Model	Prevalence of OSA	AHI Threshold	Study n	Sensitivity (95% CI)	Specificity (95% CI)	LR^+ (95% CI)	LR^- (95% CI)
BMI; Gurubhagavatula *et al.*,[44] 2004	0.281	5	406	0.702 (0.609–0.784)	0.610 (0.551–0.666)	1.797 (1.491–2.166)	0.489 (0.364–0.658)
Clinical assessment 1; Schafer *et al.*,[46] 1997	0.702	10	114	0.688 (0.574–0.787)	0.706 (0.525–0.849)	2.338 (1.360–4.016)	0.443 (0.300–0.654)
Clinical assessment 2; Schafer *et al.*,[46] 1997	0.702	10	114	0.413 (0.304–0.528)	0.912 (0.763–0.981)	4.675 (1.538–14.210)	0.644 (0.522–0.796)
Clinical and oximetry 1; Schafer *et al.*,[46] 1997	0.702	10	114	0.413 (0.304–0.528)	0.912 (0.763–0.981)	4.675 (1.538–14.210)	0.644 (0.522–0.796)
Clinical and oximetry 2; Schafer *et al.*,[46] 1997	0.702	10	114	0.338 (0.236–0.452)	0.971 (0.847–0.999)	11.475 (1.624–81.074)	0.683 (0.578–0.807)
Clinical and oximetry; Pradhan *et al.*,[47] 1996	0.570	10	150	1.000 (0.958–1.000)	0.313 (0.202–0.441)	1.452 (1.230–1.714)	0.018 (0.001–0.296)
Clinical data model; Pradhan *et al.*,[47] 1996	0.570	10	150	1.000 (0.958–1.000)	0.188 (0.101–0.305)	1.231 (1.092–1.388)	0.03 (0.002–0.496)
Clinical decision rule; Rodsutti *et al.*,[48] 2004	0.720	5	243	1.000 (0.979–1.000)	0.471 (0.348–0.596)	1.885 (1.509–2.355)	0.006 (0.000–0.097)
Clinical score; Williams *et al.*,[49] 1991	0.606	10	33	0.750 (0.509–0.913)	1.000 (0.753–1.000)	20.667 (1.343–318.08)	0.272 (0.132–0.561)
Crocker validation; Rowley *et al.*,[50] 2000	0.670	10	370	0.839 (0.787–0.882)	0.393 (0.306–0.486)	1.383 (1.187–1.611)	0.410 (0.286–0.587)
Flemons validation; Rowley *et al.*,[50] 2000	0.670	10	370	0.759 (0.701–0.811)	0.537 (0.444–0.628)	1.640 (1.337–2.012)	0.449 (0.341–0.591)
Generalized regression neural network; Kirby *et al.*,[51] 1999	0.690	10	405	0.989 (0.969–0.998)	0.802 (0.721–0.867)	4.986 (3.509–7.083)	0.013 (0.004–0.041)
Kushida index; Jung *et al.*,[52] 2004	0.771	5	70	0.938 (0.698–0.998)	0.889 (0.774–0.958)	14.222 (2.127–95.095)	0.119 (0.055–0.255)
Kushida index; Kushida *et al.*,[33] 1997	0.847	5	300	1.000 (0.923–1.000)	0.976 (0.949–0.991)	91.604 (5.815–1443.1)	0.026 (0.012–0.055)
Linear regression model 1; Vaidya *et al.*,[53] 1996	0.730	5	309	0.947 (0.909–0.972)	0.214 (0.132–0.317)	1.205 (1.073–1.353)	0.249 (0.125–0.494)
Linear regression model 2; Vaidya *et al.*,[53] 1996	0.730	5	309	0.964 (0.931–0.985)	0.226 (0.142–0.330)	1.246 (1.107–1.403)	0.157 (0.072–0.345)
MAP index bootstrapping algorithm; Gurubhagavatula *et al.*,[54] 2001	0.680	5	75	0.941 (0.838–0.988)	0.667 (0.447–0.844)	2.824 (1.597–4.992)	0.088 (0.028–0.274)
MAP index; Gurubhagavatula *et al.*,[54] 2001	0.694	5	359	0.819 (0.766–0.865)	0.700 (0.605–0.784)	2.731 (2.041–3.655)	0.258 (0.193–0.346)
MAP index; Rowley *et al.*,[50] 2000	0.670	10	370	0.871 (0.823–0.910)	0.344 (0.261–0.436)	1.328 (1.158–1.524)	0.375 (0.250–0.562)
MAP index and oximetry; Gurubhagavatula *et al.*,[44] 2004	0.281	5	406	0.746 (0.656–0.823)	0.890 (0.849–0.924)	6.804 (4.823–9.598)	0.286 (0.208–0.392)
MAP index; Gurubhagavatula *et al.*,[44] 2004	0.281	5	406	0.719 (0.627–0.799)	0.757 (0.703–0.805)	2.958 (2.344–3.733)	0.371 (0.274–0.501)
Prediction model validation; Rauscher *et al.*,[55] 1993	0.405	10	116	0.936 (0.825–0.987)	0.449 (0.329–0.574)	1.700 (1.356–2.131)	0.142 (0.046–0.438)
Prediction model; Dealberto *et al.*,[56] 1994	0.558	10	129	0.597 (0.475–0.711)	0.895 (0.785–0.960)	5.674 (2.600–12.380)	0.450 (0.335–0.605)
STOP-BANG; Chung *et al.*,[31] 2008	0.696	5	177	0.836 (0.758–0.897)	0.546 (0.423–0.679)	1.916 (1.405–2.614)	0.291 (0.183–0.462)
Symptoms; Gurubhagavatula *et al.*,[44] 2004	0.281	5	406	0.518 (0.422–0.612)	0.685 (0.628–0.738)	1.643 (1.286–2.099)	0.704 (0.574–0.865)
Viner prediction model; Viner *et al.*,[57] 1991	0.460	10	410	0.941 (0.898–0.970)	0.279 (0.221–0.343)	1.306 (1.195–1.428)	0.210 (0.114–0.386)
Viner validation; Rowley *et al.*,[50] 2000	0.670	10	370	0.959 (0.927–0.980)	0.137 (0.082–0.210)	1.112 (1.032–1.198)	0.297 (0.140–0.628)
Pooled estimates	0.281–0.847			0.835 (0.823–0.847)	0.575 (0.557–0.593)	2.003 (1.682–2.385)	0.302 (0.232–0.393)

AHI threshold = diagnostic threshold of apnea–hypopnea index used in each study; BMI = body mass index; CI = confidence interval; LR^+ = positive likelihood ratio; LR^- = negative likelihood ratio; MAP = multivariable apnea prediction; OSA = obstructive sleep apnea; STOP-BANG = acronym from Chung *et al.*[31]

test across various studies ($I^2 > 75\%$). Only the Kushida index reproducibly performed as an excellent predictor (DOR > 81) in all validated studies. A summary table of the range of FN rates was generated for each screening test (table 5), with summary recommendation for the test utility. No single questionnaire or clinical model satisfied the criteria for the ideal preoperative screening test.

Table 4. Test Characteristics of Clinical Models Predicting the Presence of Severe Obstructive Sleep Apnea

Study	Prevalence of OSA	AHI Threshold	Study n	Sensitivity (95% CI)	Specificity (95% CI)	LR^- (95% CI)	LR^+ (95% CI)
BASHIM; Dixon *et al.*,[30] 2003	0.717	30	99	0.958 (0.881–0.991)	0.714 (0.513–0.868)	0.059 (0.019–0.183)	3.352 (1.862–6.033)
BMI; Gurubhagavatula *et al.*,[44] 2004	0.281	30	406	0.772 (0.684–0.845)	0.705 (0.650–0.757)	0.323 (0.229–0.457)	2.621 (2.138–3.213)
Clinical–cephalometry; Battagel and L'Estrange,[58] 1996	0.593	25	59	1.000 (0.900–1.00)	1.000 (0.858–1.000)	0.014 (0.001–0.222)	49.306 (3.170–766.83)
MAP index; Gurubhagavatula *et al.*,[54] 2001	0.694	30	359	0.804 (0.749–0.851)	0.633 (0.535–0.723)	0.310 (0.232–0.413)	2.191 (1.699–2.825)
MAP index bootstrapping algorithm; Gurubhagavatula *et al.*,[54] 2001	0.680	30	75	0.824 (0.691–0.916)	0.958 (0.789–0.999)	0.184 (0.101–0.335)	19.765 (2.889–135.21)
Multivariable prediction; Gurubhagavatula *et al.*,[44] 2004	0.281	30	406	0.807 (0.723–0.875)	0.729 (0.675–0.780)	0.265 (0.181–0.388)	2.983 (2.421–3.675)
Multivariable prediction with oximetry; Gurubhagavatula *et al.*,[44] 2004	0.281	30	406	0.912 (0.845–0.957)	0.908 (0.868–0.938)	0.097 (0.053–0.175)	9.866 (6.857–14.195)
Predicted AI offspring; Pillar *et al.*,[59] 1994	0.361	25	105	0.316 (0.175–0.48)	0.940 (0.854–0.983)	0.728 (0.581–0.911)	5.289 (1.834–15.256)
Predicted AI validation; Pillar *et al.*,[59] 1994	0.840	25	50	0.881 (0.744–0.960)	0.250 (0.032–0.651)	0.476 (0.111–2.040)	1.175 (0.775–1.779)
STOP-BANG; Chung *et al.*,[31] 2008	0.696	30	177	1.000 (0.970–1.000)	0.364 (0.238–0.504)	0.011 (0.001–0.180)	1.571 (1.287–1.918)
Symptoms; Gurubhagavatula *et al.*,[44] 2004	0.281	30	406	0.614 (0.518–0.70)	0.620 (0.561–0.676)	0.623 (0.486–0.79)	1.615 (1.314–1.986)
Pooled estimates	0.281–0.717			0.818 (0.793–0.841)	0.732 (0.709–0.755)	0.293 (0.257–0.334)	2.755 (2.511–3.023)

AHI threshold = diagnostic threshold of apnea–hypopnea index used in each study; AI = apnea index; BMI = body mass index; CI = confidence interval; LR^+ = positive likelihood ratio; LR^- = negative likelihood ratio; MAP = multivariable apnea prediction; OSA = obstructive sleep apnea; STOP-BANG = acronym from Chung *et al*.[31]

Figures 2–5 describe the DOR characteristics of the individual questionnaires and clinical screening tests for prediction of diagnosis and severity of OSA. Although cumulative analyses were performed for all tests characteristics, the large heterogeneity precluded a more specific comment on the pooled results. Broadly, clinical models had marginally better pooled DOR than questionnaire models for both OSA diagnosis (pooled DOR 10.49 *vs.* 5.02) and severity (pooled DOR 17.24 *vs.* 10.12) prediction.

Finally, a metaregression of several study covariates and elements was undertaken to identify the source of this heterogeneity (table 6). Study characteristics with rDOR $>$ 2 were identified to be log equations (nonlinear scales), clinical models, clinical–cephalometry, combined techniques, and morphometry in ascending order of magnitude. Clinical elements associated with rDOR $>$ 2 were body mass index, hypertension, and history of choking or gasping. Covariates associated with rDOR $<$ 2 were prevalence of OSA, AHI threshold for diagnosis, publication year, number of variables, oximetry, age, witnessed apnea, neck circumference, tiredness, and daytime somnolence. History of snoring and sex balance in the study population had rDORs of 1.93 and 1.81, respectively. On diagnostic threshold analysis (table 7), the statistical significance of the regression coefficient b was found to be 0.189, suggesting that there is no influence of diagnostic threshold on diagnostic accuracy. That is, DOR is independent of the chosen diagnostic AHI threshold.

Discussion

We report the results of a meta-analysis of clinical screening tests for the prediction of diagnosis and severity of OSA. Severe OSA can be predicted by questionnaires and clinical tests with a high degree of accuracy. The Berlin questionnaire, the Sleep Disorders Questionnaire, morphometry (Kushida index), and the combined clinical–cephalometry model (Battagel) were the most accurate questionnaires and clinical models. However, there is high degree of heterogeneity and FN rate with all questionnaires and most clinical prediction models, making it possible that a significant proportion of patients with OSA will be missed by all questionnaires and most of the clinical models. Metaregression revealed that clinical models, log equations, combined techniques, cephalometry, and morphometry are significant test characteristics, whereas body mass index, history of hypertension, and nocturnal choking are significant test elements in the more accurate prediction models.

Importance of Test Accuracy: Implications for Anesthesiologists

There are several described summary measures for meta-analyses of diagnostic tests, namely sensitivity, specificity, predictive values, likelihood ratios, DOR, receiver operating characteristic curve analysis, and area under the curve. Each of these summary measures has unique advantages and disadvantages. An ideal diagnos-

Table 5. Screening Test Reliability and Summary Recommendation for Preoperative Use

Study	Pooled Study n	FN Rate	Ease of Use, 0–3	Test Accuracy by DOR	Summary Recommendation
ASA checklist	177	0.123–0.279	1	Poor	No preoperative value, unacceptable FN rate
BASHIM–Dixon	99	0.042	2	Good	No preoperative value, requires fasting insulin and HbA_{1c} levels
Berlin questionnaire	692	0.081–0.382	1	Poor–excellent	May have role in screening for severe OSA, unacceptable FN rate for diagnosis of OSA
BMI alone	406	0.228–0.298	0	Poor	No preoperative value, unacceptable FN rate
Clinical–cephalometry	59	0.0	3	Excellent	No preoperative value, requires skull and face x-ray analysis
Clinical and oximetry–Pradhan	114	0.0–0.622	3	Average–good	Requires preoperative overnight oximetry
Clinical assessment–Schafer	114	0.312–0.587	3	Average	No preoperative value, unacceptable FN rate
Clinical data model–Pradhan	150	0	3	Good	Complex methodology, may be cumbersome for routine preoperative evaluation
Clinical decision rule–Rodsutti	243	0	3	Excellent	Complex methodology, may be cumbersome for routine preoperative evaluation
Clinical score–Williams	33	0.25	3	Good	No preoperative value, unacceptable FN rate
Crocker validation	370	0.161	3	Poor	No preoperative value, unacceptable FN rate
Epworth Sleepiness scale	46	0.714	1	Poor	Unacceptable FN rate
Flemons validation	370	0.241	3	Poor	No preoperative value, unacceptable FN rate
Generalized regression neural network	405	0.011	3	Excellent	Complex methodology, may be cumbersome for routine preoperative evaluation
Kushida index	370	0.0–0.062	3	Excellent	Complex methodology, may be cumbersome for routine preoperative evaluation
Log regression model–Vaidya	309	0.036–0.053	3	Poor–average	No preoperative value, unacceptable FN rate
MAP index	1,210	0.059–0.281	3	Poor–good	No preoperative value, unacceptable FN rate
MAP index and oximetry	406	0.088–0.254	3	Good	No preoperative value, unacceptable FN rate
Predicted AI–Pillar	155	0.119–0.684	3	Poor–average	No preoperative value, unacceptable FN rate
Prediction model–Dealberto	129	0.403	3	Good	Complex methodology, may be cumbersome for routine preoperative evaluation
Prediction model–Rauscher	116	0.064	3	Good	Complex methodology, may be cumbersome for routine preoperative evaluation
SDQ–Weatherwax	242	0.149–0.25	1	Average–good	Unacceptable FN rate
Sleep questionnaire	85	0.105–0.222	1	Good	Unacceptable FN rate
Snoring questionnaire	1,409	0.696	1	Good	No preoperative value, unacceptable FN rate
STOP questionnaire	177	0.205–0.344	0	Poor	No preoperative value, unacceptable FN rate
STOP-BANG	177	0–0.164	2	Average–excellent	Excellent screening test for severe OSA, unacceptable FN rate for diagnosis of OSA
Symptoms	406	0.386–0.482	1	Poor	No preoperative value, unacceptable FN rate
Viner prediction model	780	0.041–0.059	3	Poor–average	Complex methodology, may be cumbersome for routine preoperative evaluation

Ease-of-use scale, with 0 defining easy and 3 meaning complex methodology: 1 point each for four or more test elements or variables, log scale, and need for additional techniques measurements or investigations.

AI = apnea index; ASA = American Society of Anesthesiologists; BASHIM = acronym from Dixon *et al.*[30]; BMI = body mass index; DOR = diagnostic odds ratio; FN rate = rate of missed diagnosis for any given screening test, calculated as (1 − sensitivity); Hb = hemoglobin; MAP = multivariable apnea prediction; OSA = obstructive sleep apnea; SDQ = Sleep Disorders Questionnaire; STOP = acronym from Chung *et al.*[31]; STOP-BANG = acronym from Chung *et al.*[31]

tic test in a healthy population should have a relatively high sensitivity with sufficient specificity and also be minimally intrusive, be relatively inexpensive, and identify patients early in the disease process.[17] Although high sensitivity helps to rule out OSA preoperatively, it is incomplete and potentially inaccurate as a summary statistic on its own, especially in the presence of spectrum bias and study heterogeneity. Combining sensitivity with specificity improves this, arguably at the cost of increasing complexity in terms of comparison. Although false positives could significantly increase costs directly and indirectly, because of prolonged postanesthesia care unit stay as mandated by the ASA practice guidelines[12] or more conservative local discharge policies, the bigger priority during the perioperative period is preventing mortality and morbidity related to OSA. Using FN rates as a summary measure of robustness of each screening test, we were able to describe the proportion of missed diagnosis among patients with OSA. Because sensitivity is considered prevalence independent, it follows that FN rates are also independent of prevalence of OSA. The FN rate is typically expressed as a conditional probability or a percentage. The Berlin questionnaire, which is commonly used in several hospitals now, has FN rates of 14.5–38.2%, clearly making it undependable to robustly rule out OSA preoperatively. Similar FN rates were observed with the ASA model (12.3–37.9%), STOP questionnaire (20.5–34.4%), and STOP-BANG model (0.0–16.4%) for diagnosis of OSA. FN rates tended to be marginally lower when predicting presence of severe

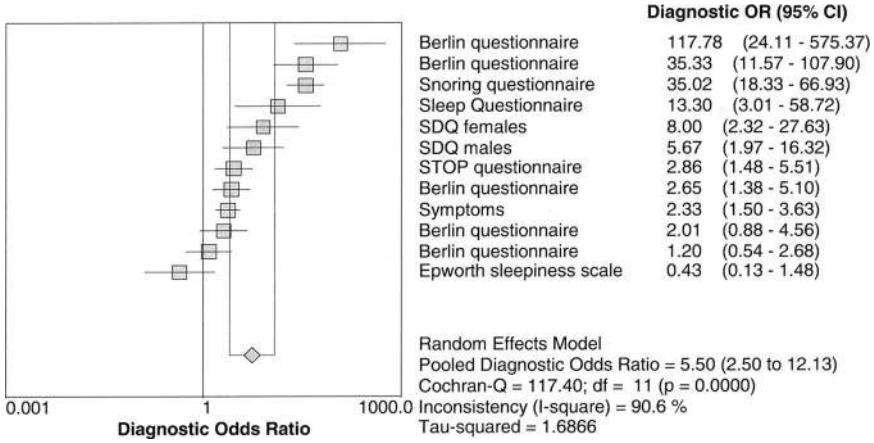

Fig. 2. Plot of diagnostic odds ratios (ORs) and 95% confidence intervals (CIs) of questionnaires predicting the diagnosis of obstructive sleep apnea. SDQ = Sleep Disorders Questionnaire; STOP = acronym from Chung *et al.*[31]

OSA across all studies. Of the remaining summary measures, positive and negative predictive values are highly dependent on disease prevalence and therefore have limited value in comparing tests. Summary receiver operating characteristic curve analysis is recommended for studies that exhibit threshold effect, but it is difficult to interpret and apply to practice. Likelihood ratios describe a user-friendly way of providing convincing diagnostic evidence ($>$ 10 positive likelihood ratio with $<$ 0.1 negative likelihood ratio) or strong diagnostic evidence ($>$ 5 positive likelihood ratio with $<$ 0.2 negative likelihood ratio). Again, the need for pairing these summary measures reduces its utility in comparative analyses. The DOR is defined as the ratio of the odds of positivity in patients with OSA relative to the odds of positivity in the nondiseased. It can also be calculated as the ratio of the positive and negative likelihood ratios and represents the best single point estimate of the receiver operating characteristic curve. Importantly, DOR used as single indicator of test performance is considered to be not prevalence dependent.[26] It provides an assessment of how well a decision tool or doctor performs in distinguishing healthy from unhealthy patients; the bigger the DOR is, the better the diagnostic accuracy is. As a result of these points, we chose the DOR as a primary measure of test accuracy. Although independent of disease prevalence, the clinical value of DOR varies with disease prevalence. A test with a DOR of 10.00 is considered to be a very good test by current standards in populations at high risk. A DOR of 10.00 in a low-risk population, on the other hand, may represent a very weak association between the experimental test and the standard test.[21] Therefore, the expectation that a highly accurate prediction tool validated for screening in sleep clinics will provide the same functionality in the preoperative period may be fallacious. One of the additional criticisms of DOR is that it ignores the relative weights of sensitivity and specificity. We accounted for this shortcoming by choosing a DOR threshold of 81, thereby ensuring that the robustness extends across both sensitivity and specificity.

The lack of reproducibility with the diagnostic performance of all the studied screening tests is an important finding of this study. Based on this study's findings, case–control studies or clinical protocols that depend on simple prediction models like the Berlin questionnaire, the ASA model, or the STOP questionnaire for assessment of high-risk groups are bound to suffer from significant FN error. Further research into predictive modeling should compare the most accurate tests in this analysis with one another in a true representative surgical population. Perhaps more importantly, anesthesiologists need to consider the importance of identifying a critical outcome measure for defining significant OSA. There is insufficient prospective evidence in the literature to support the view that mild to moderate OSA is associated with significant adverse outcome postoperatively. Indeed, regular continuous positive airway pressure therapy for mild OSA (with excessive daytime somnolence) and moderate OSA has not been shown to have the same impact on systemic disease as compared with severe OSA.[27] Further research into defining the subset

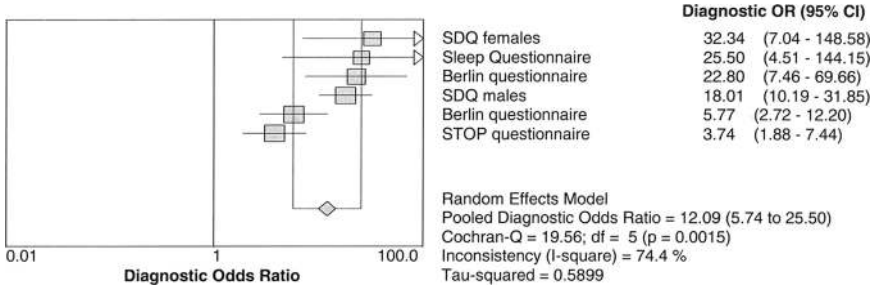

Fig. 3. Plot of diagnostic odds ratios (ORs) and 95% confidence intervals (CIs) of questionnaires predicting presence of severe obstructive sleep apnea. SDQ = Sleep Disorders Questionnaire; STOP = acronym from Chung *et al.*[31]

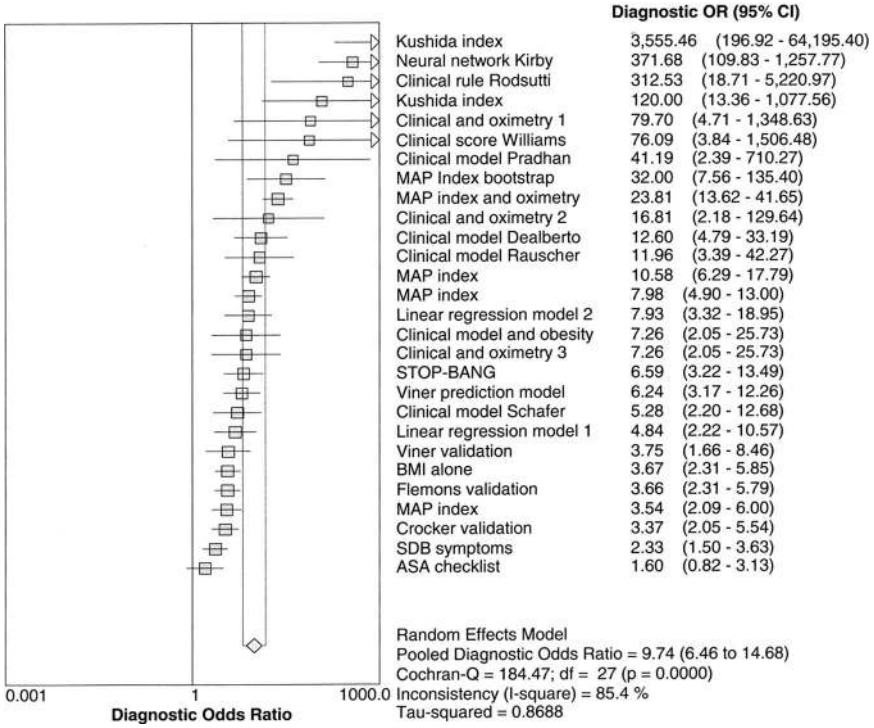

Fig. 4. Plot of diagnostic odds ratios (ORs) and 95% confidence intervals (CIs) of clinical models predicting the diagnosis of obstructive sleep apnea. ASA = American Society of Anesthesiologists; BMI = body mass index; MAP = multivariable apnea prediction; SDB = sleep-disordered breathing; STOP-BANG = acronym from Chung *et al.*[31]

of OSA patients at higher risk of perioperative mortality or morbidity is a much-needed crucial step to further avoid both unnecessary costs and potential disaster from excessive resource utilization on one hand and missed diagnoses on the other.

Analysis of Screening Test Characteristics and Elements

The characteristics of the screening tests that bestowed higher accuracy included clinical models, log equations, all combined techniques, cephalometry, and morphometry. Although questionnaires were inferior to clinical models, they are widely used currently as screening tools and therefore warrant further discussion. The Berlin questionnaire[28] was the most accurate questionnaire for predicting diagnosis of OSA. The least accurate questionnaire was the Epworth Sleepiness Scale,[29] possibly because excessive daytime sleepiness occurs commonly in obese individuals without OSA, driven by mechanisms other than nighttime sleep deprivation.[30] Regardless of the abnormal sleep physiology in obese patients, most questionnaire models have the generic problem of increasing the burden of the user for additional data collection. The simplest questionnaire, the STOP questionnaire,[31] was a poor predictor of OSA (DOR 2.9) compared with other published questionnaires. Similarly, the recently validated ASA screening tool was either of no value (DOR 1.6) or poor value (DOR 4.08) in predicting diagnosis and severity of OSA respectively. In contrast, several other clinical models trended toward significantly better performance than questionnaires in predicting diagnosis of OSA. Clinical models typically use several elements that clearly identify the OSA phenotype more robustly than questionnaires alone. A similar trend to improved accuracy was seen in clinical models *versus* questionnaires in predict-

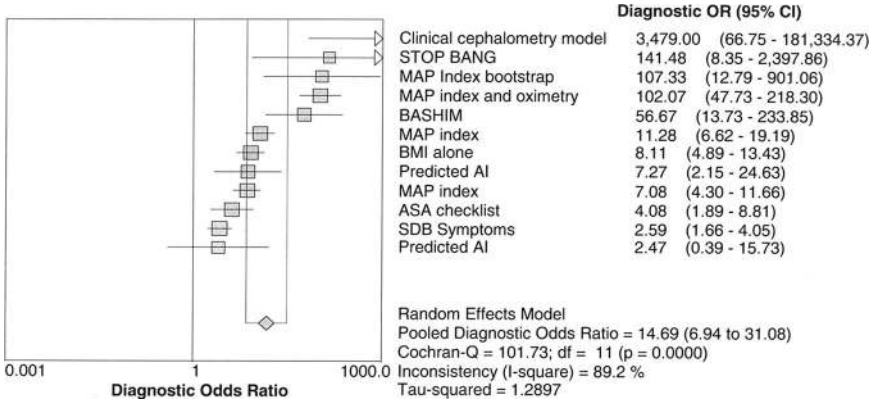

Fig. 5. Plot of diagnostic odds ratios (ORs) and 95% confidence intervals (CIs) of clinical models predicting presence of severe obstructive sleep apnea. ASA = American Society of Anesthesiologists; BASHIM = acronym from Dixon *et al.*[30]; BMI = body mass index; MAP = multivariable apnea prediction; SDB = sleep-disordered breathing; STOP-BANG = acronym from Chung *et al.*[31]

Table 6. Metaregression (Inverse Variance Weights)

Variable	Coefficient	SE	*P* Value	rDOR	(95% CI)
Study characteristics					
Oximetry	−1.545	1.0985	0.1687	0.21	(0.02–1.99)
Publication year	−1.021	0.3128	0.0025	0.36	(0.19–0.68)
Prevalence of OSA	−0.341	0.8039	0.6745	0.71	(0.14–3.64)
Study n	−0.004	0.0014	0.0068	1.00	(0.99–1.00)
Number of variables	0.056	0.0485	0.2564	1.06	(0.96–1.17)
Severity of OSA	0.303	0.1835	0.1077	1.35	(0.93–1.97)
Log equations	0.746	0.5828	0.2093	2.11	(0.64–6.89)
Clinical model	0.928	0.2854	0.0026	2.53	(1.42–4.52)
Cephalometry	2.553	2.4601	0.3067	12.84	(0.09–1,905.39)
Combined technique	2.918	1.0932	0.0116	18.50	(2.01–170.62)
Morphometry	5.659	1.2136	0.0000	286.96	(24.36–3,379.89)
Screening test elements					
Neck circumference	−0.997	0.3307	0.0048	0.37	(0.19–0.72)
Age	−0.324	0.6260	0.6085	0.72	(0.20–2.58)
Apnea	0.018	0.3754	0.9613	1.02	(0.47–2.18)
Daytime somnolence	0.325	0.4860	0.5088	1.38	(0.52–3.71)
Tiredness	0.412	0.4552	0.3723	1.51	(0.60–3.81)
Sex	0.595	0.5012	0.2434	1.81	(0.65–5.02)
Snoring	0.658	0.5013	0.1984	1.93	(0.70–5.35)
BMI	0.838	0.3891	0.0386	2.31	(1.05–5.10)
History choking or gasping	0.869	0.3766	0.0272	2.39	(1.11–5.13)
Hypertension	1.325	0.3759	0.0012	3.76	(1.75–8.08)

BMI = body mass index; CI = confidence interval; OSA = obstructive sleep apnea; rDOR = relative diagnostic odds ratio calculated by antilogarithmic transformation of covariates.

ing severe OSA. The two most accurate clinical models used additional cephalometry[32] and morphometry,[33] suggesting that using upper airway measurements could improve the accuracy of currently used clinical methods. However, the complexity of these tests could hinder their addition into standard preoperative evaluation. The STOP-BANG clinical scale[31] was identified as an excellent method for prediction of severe OSA with DOR 141.5 in one study. The ease of use of this clinical test (linear scale, and no need for additional investigations) makes it a user-friendly option for screening for severe OSA in the immediate preoperative period, although it is an average predictor of diagnosis of OSA (DOR 6.59). As with other models, further validation of this screening test is essential before final comment on its preoperative utility.

Study Limitations

Before we consider the implications of these results for clinical practice, it is important to consider some of the limitations and strengths of our methods and those of the clinical studies in our review. Although we cannot be absolutely certain that we retrieved all published material in this area, we are confident that our methodology allowed for the most thorough review of all publications from within all accessible large databases. This risk was minimized by two independent searches by the two authors. Aside from these points, there were several shortcomings in the reviewed publications, namely threshold variability, study heterogeneity, verification bias, and spectrum bias. Threshold variability refers to the influence of variable AHI threshold used explicitly or implicitly by the reference test for the validation process. Indeed, based on pooled DOR values, clinical models and questionnaires seemed to be more accurate when predicting severe OSA (AHI threshold 25 or 30). It appeared on first glance that there was sufficient variability in the chosen threshold AHI for making a diagnosis of OSA, to explain at least some of the heterogeneity seen in this meta-analysis. However, on metaregression, AHI diagnostic threshold was not seen to be a significant contributor to study accuracy in comparison with other variables. The use of a threshold AHI of 5 per hour assumes that all patients with AHI $<$ 5 per hour are morphologically and physiologically distinct from those with AHI $>$ 5 per hour. This also places mild, moderate, and severe OSA patients in one group, thereby assuming that they carry similar traits. Similarly, a threshold of AHI $>$ 30 per hour for defining severe OSA also assumes that normal, mild OSA, and moderate OSA are one homogenous group with traits uniquely different from those

Table 7. Analysis of Diagnostic Threshold

Variable	Coefficient	SE	T	*P* Value
a	2.258	0.191	11.838	0.0000
b(1)	0.106	0.080	1.329	0.1893

Moses' weighted regression (inverse variance) model (D = a + bS), where D is the natural logarithm of the diagnostic odds ratio and S is the natural logarithm of the product of the odds of true-positive test results and the odds of false-positive test results. The statistical significance of the regression coefficient b was tested to assess whether diagnostic accuracy varies significantly with changes in threshold. No evidence of threshold effect was seen.

with severe OSA. Both these assumptions are probably fallacious, because OSA encompasses a spectrum of physical and physiologic traits that overlap with normal patients. There is also some evidence that AHI < 5 per hour does not necessarily mean that the patient does not have OSA, because repeated sleep studies on consecutive days have shown discordance in diagnosis.[34] First-night effect describes the variance in AHI between the first and second nights of polysomnography, independent of duration of sleep time. There are several plausible reasons for first-night effect, including AHI, anxiety, presence of psychiatric disorders, psychoactive medications, alcohol intake, and age of the patient. The percentage of patients misdiagnosed based purely on a single-night study has been reported to be as high as 43%,[35] but these effects are seen exclusively in the mild end of the OSA spectrum. Subsequent studies have shown a significantly lower misdiagnosis rate than described above, and the American Academy of Sleep Medicine and the ASA currently recognize single-night testing as the standard for diagnosis of sleep apnea. All of these factors may explain the significant intratest heterogeneity seen with the three tests specifically analyzed for reproducibility, namely the Berlin questionnaire, multivariable apnea prediction, and the Kushida index. Any intratest heterogeneity that results in FNs is a clinically important problem for anesthesiologists, because patients with OSA may be exposed to harm. Of the studied models, only the Kushida index was deemed to be an excellent test in repeated studies.

Two main biases observed in this meta-analysis deserve further mention, namely verification bias and spectrum effect or spectrum bias. To avoid verification bias, it is important that diagnostic accuracy is assessed in consecutive patients who present with the clinical problem of interest. Clearly, all of the analyzed studies in this meta-analysis did not meet these standards. There are two types of verification bias: "partial verification," which occurs when the reference standard was not applied to all participating patients, and "differential verification," which occurs when results from a decision tool influence the choice of reference standards to apply. Estimates of sensitivity tend to be overestimated when partial verification bias is present. Both sensitivity and specificity tend to be overstated when differential verification bias is present.[36] One critical consideration common to most of the studied OSA prediction questionnaires and models is the high pretest probability of OSA, *i.e.*, all of the study patients typically attended a sleep clinic for suspected OSA or other sleep-related breathing disorders. This is a very different clinical scenario compared with the typical surgical population, where the clinical distinction between patients with and without OSA is possibly more apparent. In essence, these could be considered as two distinct study populations. Prediction models that are derived and validated in high-risk populations are subject to spectrum effect or spectrum bias, and these tests report higher sensitivity than is seen when the test is used in a lower-risk population. The advantage of deriving the screening tests in a representative healthy population is that this is exactly how the tests will be used in practice. The disadvantage is that the absolute frequency of abnormalities is much lower, meaning that confidence intervals for the results are wider unless the study has far more patients.[36,37]

In summary, this review provides a comprehensive and up-to-date synthesis of the literature regarding the accuracy of clinical screening methods in the diagnosis of OSA. It is possible to predict severe OSA with a high degree of accuracy by clinical methods that could be used preoperatively, but no single prediction tool functions as an ideal preoperative test. The Berlin questionnaire and the Sleep Disorders Questionnaire were the two most accurate questionnaires, whereas morphometry and combined clinical–cephalometry were the most accurate clinical models. However, test accuracy, as defined by DOR, has poor reproducibility in multiple validation studies of the same screening tool. Based on FN rates and heterogeneity, it is possible that all of the studied questionnaires and most of the clinical models will not correctly identify a significant proportion of patients with OSA. Because of significant differences between the validation study patients and surgical patients, further validation of the most accurate screening tests as defined in this meta-analysis is essential in a typical surgical population to identify the best preoperative method of screening for OSA.

The authors thank Ronald D. Chervin, M.D. (Professor, Department of Neurology, University of Michigan, Ann Arbor, Michigan), for his help with questions during the study and overall support of the authors' endeavor.

References

1. Young T, Palta M, Dempsey J, Skatrud J, Weber S, Badr S: The occurrence of sleep-disordered breathing among middle-aged adults. N Engl J Med 1993; 328:1230–5

2. Kaw RR, Golish JJ, Ghamande SS, Burgess RR, Foldvary NN, Walker EE: Incremental risk of obstructive sleep apnea on cardiac surgical outcomes. J Cardiovasc Surg (Torino) 2006; 47:683–9

3. Gupta RRM, Parvizi JJ, Hanssen AAD, Gay PPC: Postoperative complications in patients with obstructive sleep apnea syndrome undergoing hip or knee replacement: A case-control study. Mayo Clin Proc 2001; 76:897–905

4. Peppard PE, Young T, Palta M, Skatrud J: Prospective study of the association between sleep-disordered breathing and hypertension. N Engl J Med 2000; 342:1378–84

5. Peker Y, Kraiczi H, Hedner J, Loth S, Johansson A, Bende M: An independent association between obstructive sleep apnoea and coronary artery disease. Eur Respir J 1999; 14:179–84

6. Schafer H, Koehler U, Ewig S, Hasper E, Tasci S, Luderitz B: Obstructive sleep apnea as a risk marker in coronary artery disease. Cardiology 1999; 92: 79–84

7. Arzt M, Young T, Finn L, Skatrud JB, Bradley TD: Association of sleep-disordered breathing and the occurrence of stroke. Am J Respir Crit Care Med 2005; 172:1447–51

8. Krieger J, Sforza E, Apprill M, Lampert E, Weitzenblum E, Ratomaharo J: Pulmonary hypertension, hypoxemia, and hypercapnia in obstructive sleep apnea patients. Chest 1989; 96:729–37

9. Gami AS, Howard DE, Olson EJ, Somers VK: Day-night pattern of sudden death in obstructive sleep apnea. N Engl J Med 2005; 352:1206–14

10. Ambrosetti M, Lucioni A, Ageno W, Conti S, Neri M: Is venous thrombo-

embolism more frequent in patients with obstructive sleep apnea syndrome? J Thromb Haemost 2004; 2:1858–60

11. Flemons WW, Douglas NJ, Kuna ST, Rodenstein DO, Wheatley J: Access to diagnosis and treatment of patients with suspected sleep apnea. Am J Respir Crit Care Med 2004; 169:668–72

12. Gross JB, Bachenberg KL, Benumof JL, Caplan RA, Connis RT, Cote CJ, Nickinovich DG, Prachand V, Ward DS, Weaver EM, Ydens L, Yu S: Practice guidelines for the perioperative management of patients with obstructive sleep apnea: A report by the American Society of Anesthesiologists Task Force on Perioperative Management of Patients with Obstructive Sleep Apnea. ANESTHESIOLOGY 2006; 104:1081–93

13. Bliwise DL, Nekich JC, Dement WC: Relative validity of self-reported snoring as a symptom of sleep apnea in a sleep clinic population. Chest 1991; 99:600–8

14. Chervin RD, Murman DL, Malow BA, Totten V: Cost-utility of three approaches to the diagnosis of sleep apnea: Polysomnography, home testing, and empirical therapy. Ann Intern Med 1999; 130:496–505

15. Littner MR: Portable monitoring in the diagnosis of the obstructive sleep apnea syndrome. Semin Respir Crit Care Med 2005; 26:56–67

16. Young T, Evans L, Finn L, Palta M: Estimation of the clinically diagnosed proportion of sleep apnea syndrome in middle-aged men and women. Sleep 1997; 20:705–6

17. Ross SD, Sheinhait IA, Harrison KJ, Kvasz M, Connelly JE, Shea SA, Allen IE: Systematic review and meta-analysis of the literature regarding the diagnosis of sleep apnea. Sleep 2000; 23:519–32

18. Irwig L, Tosteson ANA, Gatsonis C, Lau J, Colditz G, Chalmers TC, Mosteller F: Guidelines for meta-analyses evaluating diagnostic tests. Ann Intern Med 1994; 120:667–76

19. Whiting P, Rutjes A, Reitsma J, Bossuyt P, Kleijnen J: The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003; 3:1–13

20. Deeks JJ: Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001; 323:157–62

21. Blackman NJ-M, ter Riet G, Kessels AGH, Bachmann LM: Systematic reviews of evaluations of diagnostic and screening tests (letter). BMJ 2001; 323: 1188

22. Zamora J, Abraira V, Muriel A, Khan K, Coomarasamy A: Meta-DiSc: A software for meta-analysis of test accuracy data. BMC Med Res Methodol 2006; 6:1–12

23. Higgins JP, Thompson SG, Deeks JJ, Altman DG: Measuring inconsistency in meta-analyses. BMJ 2003; 327:557–60

24. Lijmer JG, Bossuyt PM, Heisterkamp SH: Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med 2002; 21:1525–37

25. Littenberg B, Moses LE: Estimating diagnostic accuracy from multiple conflicting reports: A new meta-analytic method. Med Decis Making 1993; 13: 313–21

26. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PMM: The diagnostic odds ratio: A single indicator of test performance. J Clin Epidemiol 2003; 56: 1129–35

27. Barnes M, Houston D, Worsnop CJ, Neill AM, Mykytyn IJ, Kay A, Trinder J, Saunders NA, Douglas McEvoy R, Pierce RJ: A randomized controlled trial of continuous positive airway pressure in mild obstructive sleep apnea. Am J Respir Crit Care Med 2002; 165:773–80

28. 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; 131:485–91

29. Johns MW: Daytime sleepiness, snoring, and obstructive sleep apnea: The Epworth Sleepiness Scale. Chest 1993; 103:30–6

30. Dixon JB, Schachter LM, O'Brien PE: Predicting sleep apnea and excessive day sleepiness in the severely obese: Indicators for polysomnography. Chest 2003; 123:1134–41

31. Chung F, Yegneswaran B, Liao P, Chung SA, Vairavanathan S, Islam S, Khajehdehi A, Shapiro CM: STOP questionnaire: A tool to screen patients for obstructive sleep apnea. ANESTHESIOLOGY 2008; 108:812–21

32. Battagel JM, Johal A, Kotecha B: A cephalometric comparison of subjects with snoring and obstructive sleep apnoea. Eur J Orthod 2000; 22:353–65

33. Kushida CA, Efron B, Guilleminault C: A predictive morphometric model for the obstructive sleep apnea syndrome. Ann Intern Med 1997; 127:581–7

34. Wittig RM, Romaker A, Zorick FJ, Roehrs TA, Conway WA, Roth T: Night-to-night consistency of apneas during sleep. Am Rev Respir Dis 1984; 129:244–6

35. Mosko SS, Dickel MJ, Ashurst J: Night-to-night variability in sleep apnea and sleep-related periodic leg movements in the elderly. Sleep 1988; 11:340–8

36. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J: Sources of variation and bias in studies of diagnostic accuracy: A systematic review. Ann Intern Med 2004; 140:189–202

37. Mulherin SA, Miller WC: Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med 2002; 137:598–602

38. Chung F, Yegneswaran B, Liao P, Chung SA, Vairavanathan S, Islam S, Khajehdehi A, Shapiro CM: Validation of the Berlin questionnaire and American Society of Anesthesiologists checklist as screening tools for obstructive sleep apnea in surgical patients. ANESTHESIOLOGY 2008; 108:822–30

39. Ahmadi N, Chung SA, Gibbs A, Shapiro CM: The Berlin questionnaire for sleep apnea in a sleep clinic population: Relationship to polysomnographic measurement of respiratory disturbance. Sleep Breath 2008; 12:39–45

40. Sharma SK, Vasudev C, Sinha S, Banga A, Pandey RM, Handa KK: Validation of the modified Berlin questionnaire to identify patients at risk for the obstructive sleep apnoea syndrome. Indian J Med Res 2006; 124:281–90

41. Osman EZ, Osborne J, Hill PD, Lee BW: The Epworth Sleepiness Scale: Can it be used for sleep apnoea screening among snorers? Clin Otolaryngol Allied Sci 1999; 24:239–41

42. Haraldsson PO, Carenfelt C, Knutsson E, Persson HE, Rinder J: Preliminary report: Validity of symptom analysis and daytime polysomnography in diagnosis of sleep apnea. Sleep 1992; 15:261–3

43. Weatherwax KJ, Lin X, Marzec ML, Malow BA: Obstructive sleep apnea in epilepsy patients: The Sleep Apnea scale of the Sleep Disorders Questionnaire (SA-SDQ) is a useful screening instrument for obstructive sleep apnea in a disease-specific population. Sleep Med 2003; 4:517–21

44. Gurubhagavatula I, Maislin G, Nkwuo JE, Pack AI: Occupational screening for obstructive sleep apnea in commercial drivers. Am J Respir Crit Care Med 2004; 170:371–6

45. Douglass AB, Bornstein R, Nino-Murcia G, Keenan S, Miles L, Zarcone VP Jr, Guilleminault C, Dement WC: The Sleep Disorders Questionnaire, I: Creation and multivariate structure of SDQ. Sleep 1994; 17:160–7

46. Schafer H, Ewig S, Hasper E, Luderitz B: Predictive diagnostic value of clinical assessment and nonlaboratory monitoring system recordings in patients with symptoms suggestive of obstructive sleep apnea syndrome. Respiration 1997; 64:194–9

47. Pradhan PS, Gliklich RE, Winkelman J: Screening for obstructive sleep apnea in patients presenting for snoring surgery. Laryngoscope 1996; 106: 1393–7

48. Rodsutti J, Hensley M, Thakkinstian A, D'Este C, Attia J: A clinical decision rule to prioritize polysomnography in patients with suspected sleep apnea. Sleep 2004; 27:694–9

49. Williams AJ, Yu G, Santiago S, Stein M: Screening for sleep apnea using pulse oximetry and a clinical score. Chest 1991; 100:631–5

50. Rowley JA, Aboussouan LS, Badr MS: The use of clinical prediction formulas in the evaluation of obstructive sleep apnea. Sleep 2000; 23:929–38

51. Kirby SD, Eng P, Danter W, George CF, Francovic T, Ruby RR, Ferguson KA: Neural network prediction of obstructive sleep apnea from clinical criteria. Chest 1999; 116:409–15

52. Jung DG, Cho HY, Grunstein RR, Yee B: Predictive value of Kushida index and acoustic pharyngometry for the evaluation of upper airway in subjects with or without obstructive sleep apnea. J Korean Med Sci 2004; 19:662–7

53. Vaidya AM, Petruzzelli GJ, Walker RP, McGee D, Gopalsami C: Identifying obstructive sleep apnea in patients presenting for laser-assisted uvulopalatoplasty. Laryngoscope 1996; 106:431–7

54. Gurubhagavatula I, Maislin G, Pack AI: An algorithm to stratify sleep apnea risk in a sleep disorders clinic population. Am J Respir Crit Care Med 2001; 164:1904–9

55. Rauscher H, Popp W, Zwick H: Model for investigating snorers with suspected sleep apnoea. Thorax 1993; 48:275–9

56. Dealberto MJ, Ferber C, Garma L, Lemoine P, Alperovitch A: Factors related to sleep apnea syndrome in sleep clinic patients. Chest 1994; 105:1753–8

57. Viner S, Szalai JP, Hoffstein V: Are history and physical examination a good screening test for sleep apnea? Ann Intern Med 1991; 115:356–9

58. Battagel JM, L'Estrange PR: The cephalometric morphology of patients with obstructive sleep apnoea (OSA). Eur J Orthod 1996; 18:557–69

59. Pillar G, Peled N, Katz N, Lavie P: Predictive value of specific risk factors, symptoms and signs, in diagnosing obstructive sleep apnoea and its severity. J Sleep Res 1994; 3:241–4