Systematic Review and Meta-analysis of the Literature Regarding the Diagnosis of Sleep Apnea

Susan D. Ross MD, FRCPC,[1] Iris A. Sheinhait MA,[1] Katherine J. Harrison BA,[1] Marion Kvasz MD, MPH,[1] Janet E. Connelly BS,[1] Steven A. Shea PhD,[2] and I. Elaine Allen PhD[1]

[1]MetaWorks Inc., Medford, MA, [2]Harvard Medical School, and Sleep Disorders Program, Circadian, Neuroendocrine and Sleep Disorders Section, Brigham & Women's Hospital, Boston, MA

Abstract: To establish the evidence base for the diagnosis of sleep apnea (SA) in adult patients, a systematic review of the literature from 1980 through November 1, 1997 was performed. Diagnostic studies were included if they reported results of any test to establish or support a diagnosis of SA, in comparison to a diagnosis from a full polysomnogram (PSG). Test results were meta-analyzed using fixed effects models and summary receiver operating characteristic curves (ROCs) to examine consistency of tests within and between diagnostics vs. the "gold standard" of PSG.

From a total of 937 studies, 249 fit the broad eligibility criteria for inclusion in the clinical trial database and its data were extracted from these reports; useable data for statistical analyses were reported in 71 studies (7,572 patients). The sensitivity and specificity of partial channel and partial time PSGs appeared most promising as replacements for full PSG in patients suspected of obstructive SA. Clinical prediction rules (multivariate models) were also promising. Studies of portable sleep monitors, radiologic or morphologic features, and focused questionnaires were too heterogeneous to be meta-analyzed.

In general, the diversity of study designs and objectives were very high and the methodological rigor of these studies as assessments of diagnostic tests was very low. Thus, we are still not in a position to recommend standardization of diagnostic methodology for sleep apnea. Instead, our recommendations for future research include standardization of terms and diagnostic criteria, and consistently reported statistics to enhance the utility of this literature.

Key words: Diagnosis; meta-analysis; polysomnogram; sleep apnea; systematic review

INTRODUCTION

SLEEP APNEA (SA) IN ADULT PATIENTS IS A DISORDER CHARACTERIZED BY RECURRENT APNEIC AND HYPOPNEIC EPISODES DURING SLEEP. *Apnea* is usually defined as complete cessation of airflow; *hypopnea* is usually defined as a 50% or greater reduction in airflow, with or without a coincident O_2 desaturation. Most cases are characterized by recurrent airway obstruction (obstructive SA), and a minority of cases are purely central in origin. In view of its purportedly high prevalence and serious associated morbidity, SA has recently been described as a major public health concern.[1,2,3] The National Commission on Sleep Disorders Research[4] estimated that SA may be responsible for 38,000 cardiovascular deaths per year and annual costs of $42 million for related hospitalizations. The cumulative eight-year mortality of untreated SA has been estimated as high as 37% for patients with an apnea index (AI) $\geq$20, where apnea index is defined as the number of apneic episodes/hour sleep, compared to 4% for patients with lower AIs.[5] Patients with obstructive SA need not go untreated, however, since there is a well-established first line therapy — continuous positive airway pressure (CPAP). A major problem in the field, however, is diagnosis: who to test, how to test, and what are the implications of test results regarding the risk of serious clinical sequelae?

SA is a condition where the widely accepted standard diagnostic method (overnight full polysomnography [PSG] attended by trained personnel in a sleep laboratory) is intrusive and costly, and the interpretation can be difficult. A standard PSG typically consists of at least two channels of electroencephalogram, submental (± tibialis) electromyogram, two channels of electrooculogram, respiratory airflow, respiratory effort (thoracic and abdominal breathing movements), oxygen saturation (oximetry), and electrocardiography. Body position and snoring (microphone) are also frequently monitored in formal sleep studies. The contribution of each of these components to the PSG diagnosis of SA has not been well substantiated.[6] Even these widely accepted diagnostic techniques are still evolving. For example, the recent introduction of nasal pressure recording is replacing the respiratory airflow measurement

Accepted for publication December 1999

Address correspondence to: Iris A. Sheinhait, MetaWorks Inc., 10 Presidents Landing, Medford, MA 02155, Telephone: 781-395-0700, Fax: 781-395-7336, E-mail: Iris@MetaWork.com

because in combination with respiratory effort, nasal pressure is much more sensitive in detection of obstructive hypopnea.[7]

If the estimated prevalence of SA at 2% to 4% percent of middle-age adults is correct,[2] the cost of full PSGs for all suspected cases would be prohibitive. However, recent research on the cost-effectiveness and economic implications of diagnostics for sleep apnea showed that full PSG was the most cost-effective of PSG, home systems, and empirical therapy.[8] The development of simpler and less costly alternatives for diagnosis or pre-PSG screening is highly desirable. Diagnostic approaches that might be viewed as alternatives to PSGs or as screening tests to better select patients for PSG include: partial channel PSGs, partial night or daytime PSGs, portable sleep monitoring devices for use at home, radiologic imaging of the head and neck for anatomic abnormalities predictive of SA (including cephalometry, MRI and CT scans), anthropomorphic measurements (such as neck circumference, nasopharyngeal and laryngeal endoscopic measurements of upper airway structure and function), and focused questionnaires.

SA can be viewed as an “orphan condition,” shared by many healthcare specialties yet owned by none. Neurology, psychiatry, dentistry, otolaryngology, pulmonology, and internal medicine all share diagnosis and management of SA, and as a result, the evidence base is uneven and dispersed, and clinical management perspectives are sometimes in conflict. When evidence is scattered, and possibly conflicting, a rigorous and comprehensive assessment of all of the best available evidence is critically important and, in the case of SA, long overdue. Therefore, the aim of the current study was to develop an evidence base relevant to answering the following key questions concerning the diagnosis of SA: 1) What diagnostic and screening tests are presently available 2) What is the strength of the evidence in support of each? The analysis was designed to evaluate the diagnostic accuracy of alternatives to full PSG for the diagnosis of sleep apnea as compared to full PSG. This evidence base was developed via a systematic review of the sleep apnea literature published in the five major Western European languages. This evidence base, if kept updated, should be useful in the development of evidence-based strategies and algorithms to guide the diagnostic work-up of patients suspected of SA. This work should provide guidance for future researchers to generate new data to fill the information gaps discovered during the review. The following is a report of the methods and chief findings of this systematic review.

Methodology

In general, we used state-of-the-art systematic review methods derived from the evolving science of review research.[9,10,11] It was not our intent to review technical considerations of various tests and devices. Readers are referred to the American Academy of Sleep Medicine 1994 statement on portable devices for discussion of technical issues related to data acquisition, storage, retrieval, and analysis.[12] Our review followed a prospective protocol designed pre-data extraction, which outlined the methods to be used for the literature search, study eligibility criteria, data elements for extraction, and methodological strategies to minimize bias and maximize precision during the process of data collection, extraction, and synthesis. The protocol was shared with a panel of four experts in the field of SA prior to implementation, and the final report was shared with a panel of 15 experts in the field of SA, and thus incorporated input from representatives of insurers, government personnel, medical specialty societies, sleep disorder associations, and consumer groups. Review by members of an organization does not imply endorsement by that organization.

Literature Search

The published literature was searched from 1980 through November 1, 1997 and the retrieval cut-off date

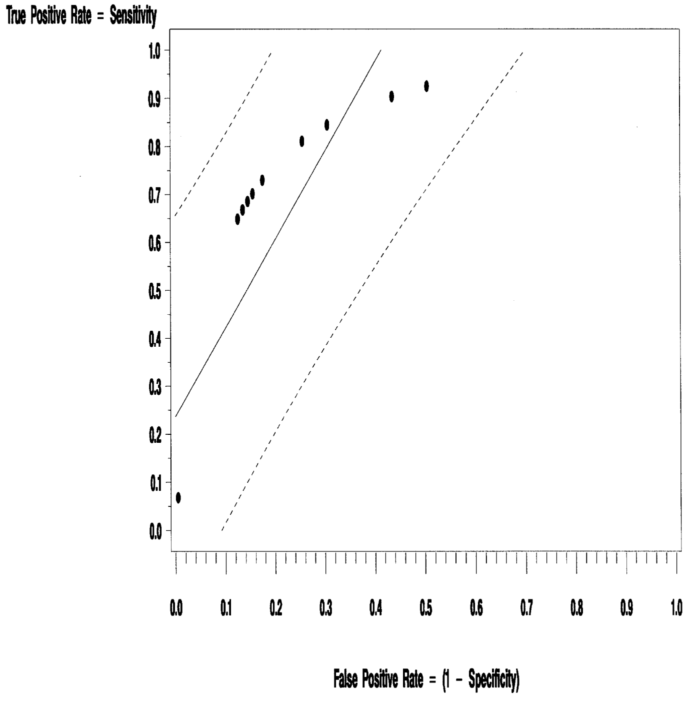
Table 1—Summary of study patient level characteristics by type of test

Study sets	K*	Mean evidence score (range)	# of patients	% males (k)	Mean age (k)	Mean BMI (kg/m^2)	Sensitivity (%)	Specificity (%)
Total analyzable studies	71	20.6 (16-34)	7,572	81 (61)	49.0 ± 5.1 (65)	30.9 ± 3 (44)	ND	ND
Partial time PSG	7	18.6 (17-20)	505	86 (6)	51.4 ± 3.5 (7)	33.9 ± 5.5 (5)	69.7 ± 5.3	87.4 ± 5.4
Partial channel PSG	3	17.7 (17-19)	213	81 (3)	51.7 ± 1.2 (3)	32.0 ± 1.0 (3)	ND	ND
Oximetry	12	20.0 (16-32)	1,784	83 (10)	50.6 ± 4.7 (11)	31.7 ± 1.1 (12)	87.4 ± 3.8	64.9 ± 6.7
Portable devices	25	22.1 (16-34)	1,631	84 (21)	48.5 ± 5.5 (24)	30.0 ± 1.6 (15)	ND	ND
Prediction equations	8	21.5 (17-30)	1,908	77 (8)	49.4 ± 4.0 (8)	31.4 ± 3.4 (8)	66.5 ± 14.0	88.7 ± 4.9
Flow volume loop	4	18.3 (17-20)	594	79 (4)	50.0 ± 1.6 (4)	29.0 (1)	39.1 ± 25.3	60.5 ± 23.7
Global impression	4	23.3 (19-28)	1,139	67 (4)	47.7 ± 2.1 (3)	29.4 ± 0.7 (3)	58.9 ± 4.2	65.6 ± 4.8
Other clinical	9	19.8 (18-22)	815	91 (8)	47.4 ± 6.9 (8)	30.3 ± 1.5 (6)	ND	ND
Chemical	1	18	88	49	58	28	ND	ND
Radiologic	5	18.5 (17-20)	296	73 (3)	43.0 ± 4.7 (4)	30.5 ± 4.4 (4)	ND	ND
Questionnaire	3	19.0 (17-21)	576	58 (2)	45.3 ± 4.7 (2)	28.0 (1)	ND	ND

*Eight studies (#8A- Garcia-Diaz, et al., 1997, #18A- Gugger, 1997, #19A-Gyulay, et al., 1993, #46A- Hoffstein and Szalai, 1993, #-49A-Pracharktam, et al., 1996, #44A- Schafer, et al., 1997, #20A- Svanborg, et al., 1990, #45A-Viner, et al., 1991.) report results for more than 1 test, consequently the sum of all the, categories is greater than 71. ND=Not Done; *k=# of studies was January 30, 1998. The search started with a broad Medline search using the terms "sleep apnea syndrome," and "monitoring, physiologic," "sleep apnea syndrome," and "airway resistance," and "human." Investigators also searched "sleep apnea syndromes," "sleep apnea syndrome" and "index." In addition, the 1997 Current Contents® CD-ROM was searched ("sleep apnea") to the same cut-off date. All citations and their full papers' abstracts were printed and screened for any mention of diagnostic tests in adults with SA (Level I screening). Abstracts were rejected at Level I screening based upon information presented in their abstracts for the following reasons: (1.) treatment papers; (2.) peripheral topics; (3.) reviews; (4.) case studies; (5.) special populations of patients (e.g., patients with neuromuscular diseases or cerebral malformations, congenital or acquired structural abnormalities of the head or neck); (6.) pediatric studies. All studies passing Level I screening were retrieved for second screening (Level II). To be eligible for inclusion, studies had to enroll at least 10 adult patients with any form of SA (obstructive, central, mixed, or not specified) undergoing any diagnostic test or intervention to establish or support a diagnosis of SA in comparison with full PSG. Studies reported in the five main Western European languages (English, French, German, Italian, or Spanish) were eligible. The electronic searches noted above were supplemented by a thorough search of the reference lists of all eligible studies and relevant review articles. Relevant Internet sites posted by medical specialty societies and patient advocacy groups were contacted for identification of any additional pertinent information about current recommendations or guidelines for assessment of disease status in patients suspected of SA.

Figure 1: Partial Time PSG Tests

Summary ROC Curve and 95% CI from Meta-Analysis of Individual Studies



Rating the Evidence

All potentially eligible diagnostic studies were rated by senior investigators (two MDs, one PhD) to assess validity of each study as a diagnostic test study prior to data extraction. A customized rating instrument was used, derived from 1) the assessment guide provided by Irwig et al.[13] for assessing validity of studies of diagnostic tests in general, and 2) features important to SA studies in particular, as suggested by Flemons and Remmers.[14] In general, studies that received the highest scores used full PSG results as the "gold standard" against which a second test was evaluated, with random order assignment of tests and PSG, and with blinding of the readers of each test to the results of the other tests. Several other features of diagnostic study design, execution, and reporting were also rated including types of outcomes evaluated and statistical tests performed. Possible scores ranged from 0 to 44, with higher scores suggesting higher quality of diagnostic test evidence. Papers scoring less than 16 points (i.e., falling in the lowest 20% of the distribution of actual scores) were dropped from further consideration for data extraction and analysis. This was a post-hoc decision made upon recognition that low scoring studies did not generally have analyzable diagnostic test results.

Data Extraction and Database Development

Data from each study were extracted in duplicate by investigators using data extraction forms developed and tested for this review, with one extractor using a blinded copy of each study report (masked as to source of financial support, authors, and journal). The data extraction forms, completed independently by the two investigators, were then compared, and differences were resolved by consensus, referring to the information in the original report as necessary. Only clearly reported aggregate results were extracted from studies. Results that were only given for individual patients, and results that would require extrapolations from equations, graphs, or derivations from figures or tables were not extracted.

Key data elements sought for extraction from each diagnostic study included study descriptors, patient demographic features, (including concomitant illnesses, signs, and symptoms of sleep apnea syndrome), and test characteristics. The following features were sought for test characteristics:

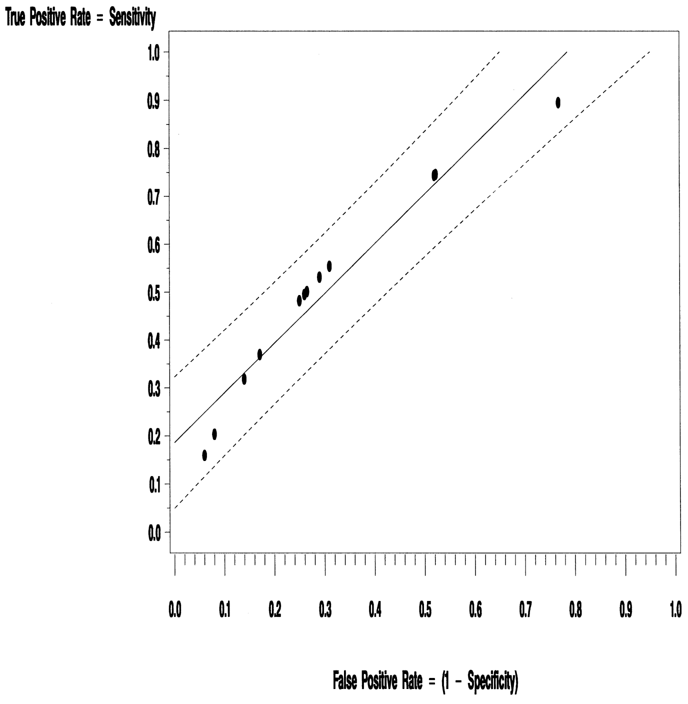
· PSG type: full vs. partial monitoring
· Full night vs. partial night vs. daytime PSG results

· Apnea index (AI) or hypopnea index (HI), or apnea-hypopnea index (AHI). AHI refers to the total apneas plus hypopneas during total time asleep, divided by the number of hours asleep. The respiratory disturbance index (RDI) is the same as AHI.
· Portable devices: test metric, thresholds for diagnosis, results, site (home vs. laboratory) and conditions (full night vs. partial night vs. daytime)
· Methods of all sleep test analyses (computer vs. manual, sleep time vs. time in bed, or test time, and definition of apnea and hypopnea episodes)
· Non-sleep tests: clinical, radiologic, chemical, questionnaires, prediction equations, etc., with test metric and thresholds for diagnosis or next action, and results
· Results of all reported statistical tests: sensitivity, specificity, positive predictive value, negative predictive value, and correlation coefficients of each test relative to full PSG results

Statistical Methods and Graphical Analysis

The main objective of the analysis was to evaluate the diagnostic accuracy of alternatives to full PSG for the diagnosis of SA as compared to a full PSG. For the analyses, PSG was used as the "gold standard". PSG was either stated to be "full" or "standard" by the authors, or included at least the following parameters: oximetry, thoracoabdominal respiratory excursions, airflow, submental electromyogram (EMG), electroencephalogram (EEG), and electrooculogram (EOG). In order to be eligible for the statistical analysis, studies had to report outcomes in terms of the sensitivity and specificity (or a function of these outcomes— i.e., likelihood ratios) of the new test as compared to the results (AI, AHI, RDI) of a standard PSG. If the sensitivity and specificity were not reported, sufficient information on the performance of the test regarding the true positive and true negative outcomes had to be reported in order to calculate sensitivity and specificity, or, in some cases, a correlation coefficient between the alternative test and the diagnosis of SA by full PSG. Correlation coefficients were extracted, but due to lack of data, they could not be analyzed.

Figure 2: Oximetry Tests
Summary ROC Curve and 95% CI from Meta-Analysis of Individual Studies



To account for the different numbers of patients in each study, weighted averages using Mantel-Haenszel fixed effects models[15] combining the comparative summary statistics, were calculated and summarized for groups based on diagnostic test category.[13] Study and patient-level covariates were also summarized for each diagnostic category, and weighted by study size when appropriate.

The statistical assessment of diagnostic tests was performed through the comparison of their sensitivity and specificity vs. the full PSG. A receiver operating characteristic curve for an individual study will display the effect of changing diagnostic cut-off values (in this case, AI or AHI) upon the sensitivity and specificity of the test. A summary ROC curve in effect combines individual study ROCs in a meta-analytic framework (weighting by study size and variance) to give an overall picture of diagnostic accuracy of a test over the range of cut-off values represented.

A summary ROC curve was calculated for each diagnostic test group where sufficient data were available.[16,17,18] The resulting curve describes how the test's performance in those with SA (sensitivity or true positive rate [TPR] varies with its performance in those without SA); (1 - specificity or false positive rate [FPR]). The summary ROC plot represents each study as a single point and the curve represents the overall summary of all studies eligible for inclusion in that specific analysis with each study weighted by study size. The 95% confidence intervals are also displayed. Where the studies give similar results, the curve and 95% confidence bound will be close to the points. All calculations were performed using SAS® software version 7.0.

RESULTS

Search Results

The initial search through Medline and Current Contents® yielded 3,730 citations. An additional 202 citations were identified from a manual search of reference lists. Most citations were rejected at Level I screening due to the following reasons: ineligible patient populations, a treatment study, or no studies of adult SA syndrome. After screening these citations, 937 studies were potentially eligible. Of these 937 studies, 249 fit the broad eligibility criteria for inclusion in the clinical trial database and data were extracted from these reports. Analysis of the extracted data yielded 71 studies that reported outcomes with sensitivity, specificity, and/or correlations, relative to a full PSG. These 71 studies were the subject of the subsequent analysis and the results are described below. All accepted 71 studies are listed in Appendix A (1A–71A) and extracted data listings may be provided to readers by the authors upon request.

Study Characteristics

Of the 71 diagnostic studies (1A–71A) reporting outcomes formats required for inclusion in the analyzable dataset (full PSG, sensitivity and specificity, or correlation coefficients), 63 were in English, 5 in German, and 3 in Spanish. Twenty-one studies were performed in the U.S., 33 in Europe, and 17 elsewhere. These studies were published from 1981 to 1997, with 52 studies published since 1991. The average diagnostic evidence score was 20.6 (range 16 to 34). The range of scores in the included data set was narrow, due to our prior rejection of low scoring studies from the analyzable data set.

In total, there were 7,572 patients enrolled, and the average number of patients per study was 106.6 (range 10 to 594). Only three stated industry sponsorship. Of the 71 studies with full PSG as the “gold standard”, there were 12 oximetry alone studies, seven partial time PSGs, three of partial channel PSGs, 25 studies with results of portable monitoring devices, 17 studies with clinical assessments (including flow volume loops and global impressions), only one study with chemical assay, five reporting radiologic test results (one MRI, three cephalometry, and one reporting CT and cephalometry), and three studies with focused questionnaires. Also, eight studies reported results of multivariate models as predictors of PSG results. Eight studies (18A–20A, 44A–46A, 49A) reported results for more than one test, consequently the sum of all the categories is greater than 71.

Patient Characteristics

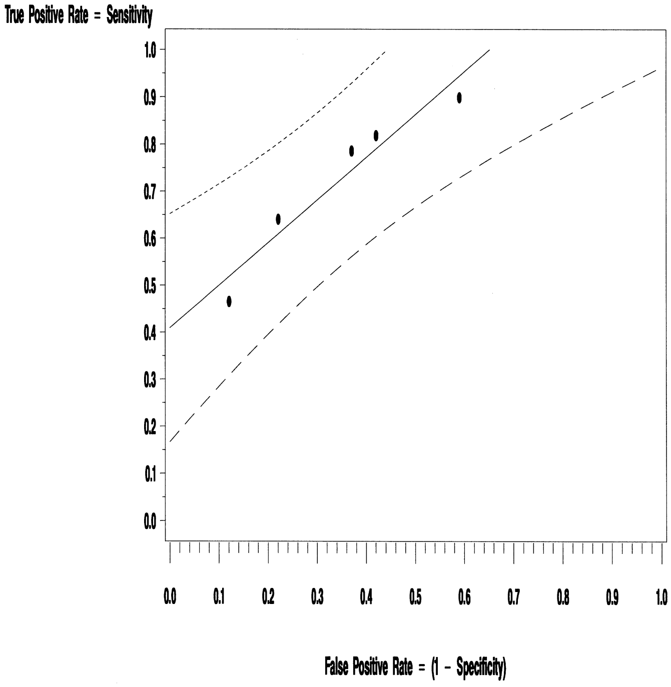
Of the 4,400 patients reported to have a suspected sleep disorder in this set of studies, a PSG diagnosis of SA was made in 2,037 (49 %), using the lowest apnea index (AI) (i.e., least specific, most sensitive) or apnea-hypopnea index (AHI) diagnostic thresholds reported. Note the PSG definitions of apnea and hypopnea and sleep apnea diagnostic cut-off criteria varied somewhat from study to study. The severity of SA was too infrequently reported to analyze this further. Mean age was 49.0±5.1 years (range 36 to 60) for the 7,572 patients studied. Of the 61 studies in this set which reported gender, 81% of patients were male. Body mass index (BMI) was reported in 44 studies, and averaged 30.9 kg/m^2 (range 26.2 to 40.0).

The relationships between patients' medical histories and PSG results, and patients' relevant symptoms and PSG results could not be estimated because patients' histories and symptoms were not consistently reported in these 71 studies.

Diagnostic Test Characteristics

All included diagnostic studies were required to report results of a 'standard' PSG. However, criteria that constitute a standard PSG varied among the studies. Most studies monitored respiratory activity (chest and abdominal movements and airflow) and oxygen saturation (oximetry). All included measures of sleep (EEG, EOG, or submental EMG), and some included measures of cardiac activity (ECG). Tibial EMG, snoring, and body position were less frequently monitored. Most noted that the traditional scoring system for sleep stages of Rechstahffen and Kales[19] was used, and PSG data were assessed manually in nearly all cases. Apnea was typically defined as complete cessation of airflow, but in some studies, a >80% reduction in airflow was used. For defining hypopnea, most papers suggested a 50% or greater reduction in airflow, with or without a coincident O_2 desaturation of anywhere from 2% to 4% from some average SaO_2 over a preceding interval of time. Nearly all studies based the AI (or AHI) upon the time asleep, as opposed to the time spent in bed, except for the portable devices intended for home use which are described

Figure 3: Portable Devices with Oximetry, HR, Snoring, and Body Position
Summary ROC Curve and 95% CI from Meta-Analysis of Individual Studies

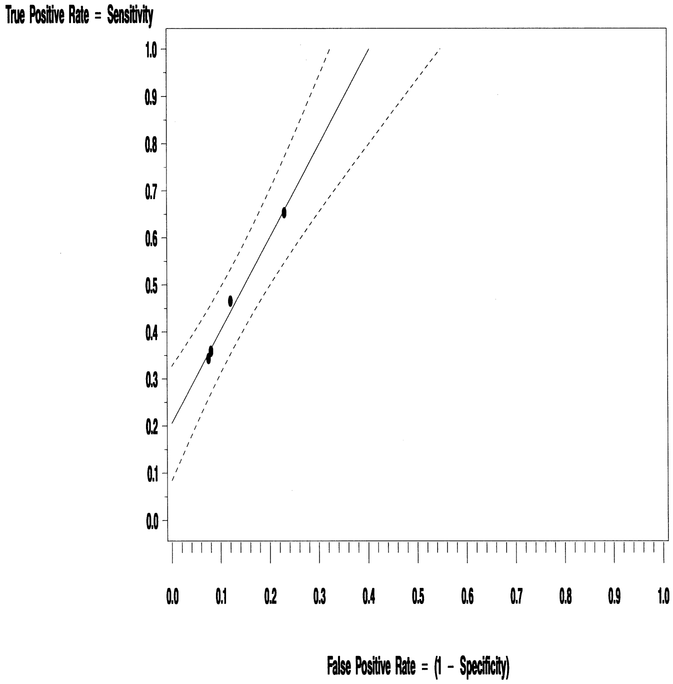


further below. All standard PSGs were performed in the sleep laboratory, which was either a free-standing unit, or a hospital setting with trained attendants present. Different thresholds for AI or AHI (or RDI) were used in different settings to make a diagnosis of SA, ranging from 5 to 15 for AI and 5 to 40 for AHI. The most frequently used cut-off was 10 for AI followed by 15 for AHI. Some studies required the presence of signs or symptoms of sleep disturbance together with an elevated AI (or AHI) for SA diagnosis, and some did not. Most did not report distinctions between obstructive SA, central SA, or mixed SA. Studies are summarized according to test (see Table 1) and described further below.

In the following sets of studies, we have applied a meta-analytic[16] method to summarize the diagnostic accuracy of the tests therein. Simply averaging the true positive rate [TPR], and the false positive rate [FPR] from each of a set of studies would be very misleading since a single point cannot show the relationship between TPR and FPR which results from varying diagnostic cut-offs. As the diagnostic cut-off varies, the balance of sensitivity and specificity shifts. The summary ROC curve can summarize multiple studies and data on the accuracy of each diagnostic test without necessarily knowing the exact diagnostic cut-off (AI or AHI) for positivity used in each of the reports. This is possible because the data already incorporate the effects of varying diagnostic thresholds. Most of the variation in diagnostic test accuracy in these reports will be derived from these threshold effects, since other potential causes of variability, the condition, patients, prevalences, and index test (PSG) are assumed to vary little in our included studies. If all the points fall near the summary ROC curve, their differences can be attributed mostly to differences in diagnostic thresholds used. If there is wide scatter for the points around the ROC curve, other factors that affect TPR and FPR (such as listed above) may well be contributing to the variation in diagnostic accuracy thus displayed.

Figure 4: Portable Devices with Airflow and Oximetry

Summary ROC Curve and 95% CI from Meta–Analysis of Individual Studies



Partial Time Polysomnogram: There were seven studies (1A–7A) reporting results with sensitivity, specificity, and/or correlations of partial night or day PSGs relative to full night, standard PSGs. The average evidence score of all seven studies was 18.6, with a narrow range, from 17–20. All PSGs were performed in sleep laboratories with the standard array of physiologic monitors. Four studies compared partial night to full night PSGs, and the other three studies compared daytime PSGs to full night PSGs. These studies included 505 patients in total, most of whom were suspected of SA. The number of patients with a diagnosis of SA was not completely reported. Their average age was 51.4 (seven studies reporting) and the percentage of patients who were male was 86% (six studies reporting). The average BMI was 33.9 kg/m^2.

Of the four studies with comparisons of partial night to full night PSGs (1A, 2A, 4A, 7A), all used AHI as the PSG metric for diagnosis of SA. One also provided results using AI. One of these studies only reported correlations, not sensitivity or specificity. Of the remaining three studies, the sensitivity of the partial night PSG ranged from 42% to 93%, and the specificity ranged from 70% to 100%. However, these ranges reflect varying AHI thresholds for diagnosis of SA.

Of the three studies of daytime PSG compared to full night PSG (3A, 5A, 6A), two used AHI and one used AI as the PSG metric for diagnosis of SA. The sensitivity of the daytime PSG for results of full night PSG ranged from 66% to 100%, and the specificity ranged from 50% to 100%, again depending upon the AHI or AI thresholds used for diagnosis.

The summary ROC curve derived from these studies is presented in Figure 1. Most studies were quite homogeneous with one exception which had low sensitivity and extremely high specificity. Sensitivity at AI/AHI threshold of five was 69.7 % (± 5.3) and improved at thresholds of 10% to 79.5% (± 5.2). Specificity at AI/AHI threshold of 5 was 87.4% (± 5.4) and at the higher threshold of 10, changed little, at 86.7% (± 4.6). At still higher AI/AHI thresholds, there were too few studies with analyzable results.

Partial Channel Polysomnogram: In three studies (8A, 9A, 10A) results of a partial set of PSG channels monitored for a full night were related to the full channel, full night PSG results. The average evidence score was 17.7

(17, 17, and 19). In all three studies, oximetry, airflow, and thoracoabdominal movement were recorded. In two studies, patients were monitored on two different nights, and in the third study, same night results were compared using respiratory channels vs. full PSG. These studies totaled 213 patients with suspected or confirmed SA. Their average age was 51.7 (3 studies reporting) and percentage of patients who were male was 81% (three studies reporting). The average BMI was 32.0 kg/m^2. Sensitivity ranged from 82 to 94 % and specificity from 82% to 100%. There were too few studies to meta-analyze.

Oximetry: Sensitivity, specificity, and/or correlation of oximetry results to standard PSG results were reported in 12 studies (8A, 11A-21A). The evidence score ranged from 16–32 out of a possible 44 points and the mean score was 20.0. In three of these studies (15A, 16A, 19A), oximetry was measured separately (different time and setting) from the PSG, including overnight at home in two (16A, 19A) and on two different nights in one (19A). In the other nine studies, oximetry was measured during the nocturnal PSG. No studies were included where results from the oximetry channel on a multi-channel portable device were compared with PSG results, as this was invariably a post hoc result. The publication dates spanned 1986 to 1997. There were 1,784 patients in total, 1,756 of whom were suspected of having SA. The number of diagnosed SA patients was not reported in all studies. Their average age was 50.6 (11 studies reporting) and percentage of patients who were male was 83% (10 studies reporting). The average BMI was 31.7 kg/m^2.

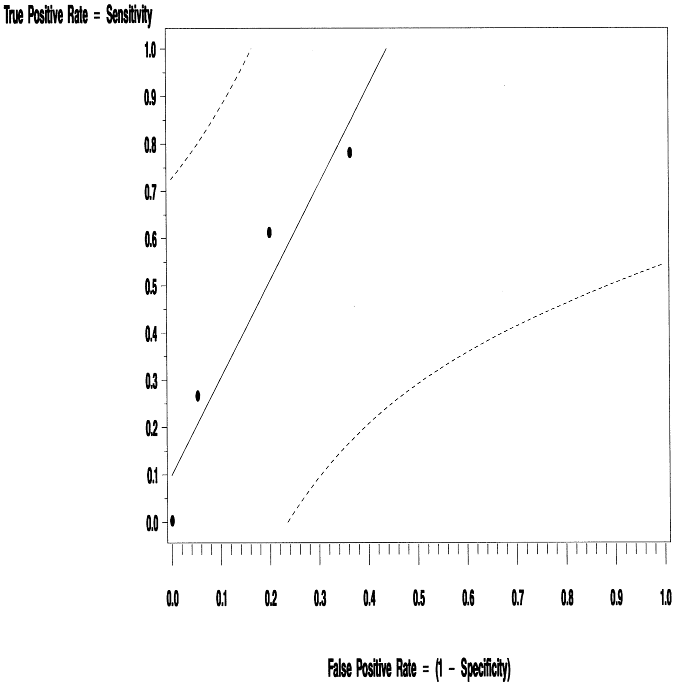
These oximetry studies used various formats for presentation of results: hourly frequency of desaturations of 3 % or 4%, frequency of desaturations less than 90 %, O_2 variability, or patients exceeding a certain number of desaturations per hour or per night. The type of probe was not consistently reported. Overall sensitivity of oximetry in individual studies ranged from 36% to 100% and specificity ranged from 23% to 99% percent, with varying AI/AHI thresholds. The overall estimates (with standard error) of sensitivity and specificity are 87.4% (± 3.8) and 64.9% (±6.7), and the summary ROC curve was generated (Figure 2). This curve shows the adjusted study sensitivity and specificity and the ROC curve calculated from the meta-analysis of these rates that the studies all fell close to the estimated curve indicating little heterogeneity.

Portable Devices: In total there were 25 portable device studies (18A, 20A, 22A–44A) with sensitivity, specificity, and/or correlation to standard PSG. The average evidence score was 22.1 (range 16–34). These studies enrolled 1,631 patients, and 1,368 were suspected SA patients at entry. Their average age was 48.5 (24 studies reporting) and percentage of patients who were male was 84% (21 studies reporting). The average BMI was 30.0 kg/m^2.

Of the 854 suspected SA patients whose subsequent diagnosis was reported, 500 (58.5%) were diagnosed with SA, using an AI/AHI threshold of $\geq$5/hr. In all studies except two (42A, 44A), the portable device results were only available as measured in the setting of a sleep laboratory, and not at home, where they are generally intended for use. Comparison of portable devices with full PSGs were therefore, in reality comparisons of partial montages vs. full montages. Differences in the settings and in the technical quality of the signals were not addressed in these studies. Devices were issued from different manufacturers, thresholds used for diagnosis of SA varied from 5 to 40 (AI or AHI) per hour, and results were reported in different formats. These devices were subdivided into groups by the channels they measured and summarized where data were available.

Five studies (25A, 33A, 38A, 39A, 44A) measured oximetry, snoring sounds, heart rate, and body position. These studies included 444 patients total. Figure 3 gives the summary ROC curve for these studies. Four studies (18A, 23A, 27A, 32A) recording airflow and oximetry enrolled a total of 178 patients. Figure 4 gives the summary ROC curve for these studies. Six studies (24A, 34A, 35A, 37A, 42A, 43A) tested portable devices monitoring oximetry, airflow, breathing and heart rate. Some of them also included measurements of body position, body movement and snoring sounds. There were 436 patients enrolled in these studies. The summary ROC curve for the four

Figure 5: Portable Devices with Airflow, Respiration, Oximetry, HR, Body Position, and Snoring
Summary ROC Curve and 95% CI from Meta-Analysis of Individual Studies



studies which included airflow, respiration, oximetry, heart rate, and body position are shown in Figure 5. The remaining eight studies could not be grouped by channels. Sensitivity of these devices ranged from 78% to 100% and specificity ranged from 62% to 99.5%.

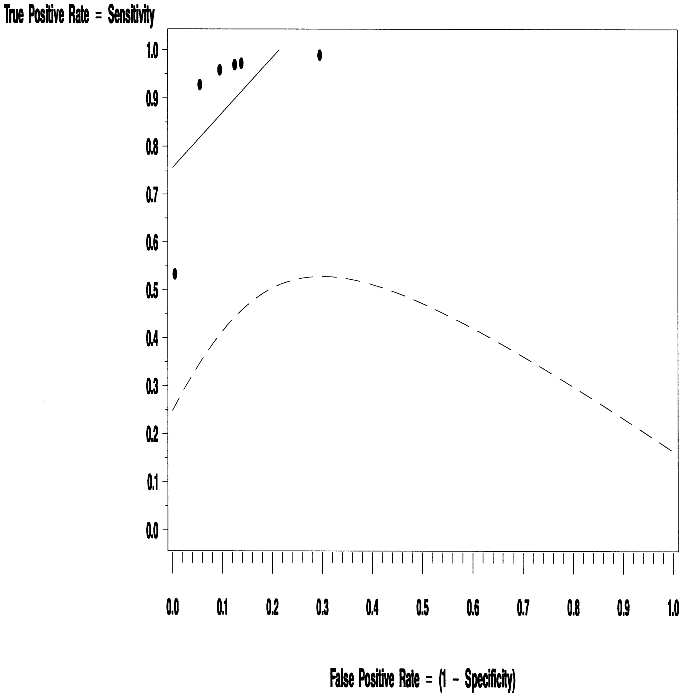
Non-sleep Tests

There were 17 studies which provided sensitivity, specificity, and or correlations of results of some clinical measure in relation to standard PSG results. Pulmonary function tests and flow volume loops were included in this set of studies.

Prediction Equations: Eight studies (19A, 44A, 45A, 46A, 47A, 48A, 49A, 50A) reported the sensitivity, specificity, or correlations of multivariate models relative to PSG results. The evidence scores of these studies ranged from 17 to 30, and averaged 21.5. The predictive accuracy of each separate component of each model was not extracted, although it was reported in some studies. Only the predictive features of the model as a composite result were captured. These studies included 1,908 patients, 254 of whom were known at entry to have SA; an additional 841 were diagnosed during the study. Their average age was 49.4 (eight studies reporting) and the percentage of patients who were male was 77% (eight studies reporting). The average BMI was 31.4 kg/m^2. Each model included at least three of the following variables: gender, age, obesity, hypertension, neck circumference, overjet (the horizontal measurement from the labial of the maxillary central incisor to the labial of the mandibular central incisor), BMI, cephalometry measurements, arterial blood gases, home oximetry, pulmonary function tests, apnea spells, snoring, falling asleep while driving, and percentage of time spent in stage 1 sleep. Sensitivity of the models for the PSG result ranged from 28% to 97.6%, and specificities ranged from 21.4% to 100%. The pooled estimate for sensitivity was 66.5% (±14.0) and specificity was 88.7 percent (±4.9). The summary ROC curve is shown in Figure 6. In this figure, both sensitivity and specificity were high.

Figure 6: Prediction Equations

Summary ROC Curve and 95% CI from Meta-Analysis of Individual Studies



Flow Volume Loops: Four studies (51A–54A) reported results of flow volume loops. These studies included 595 patients total, of which 286 were diagnosed with SA (one patient with pure central apnea (enrolled in study 54A was excluded from all analyses). The evidence score of these studies ranged from 17 to 20 (average=18.3). Their average age was 50.0 (four studies reporting) and percentage of patients who were male was 79% (four studies reporting). The average BMI was 29.0 kg/m^2. PSG results were expressed as AI in two studies (diagnostic cut-offs, 5 and 10) and AHI in one study (diagnostic cutoff 10). One study did not state the PSG metric used for SA diagnosis. Sensitivity of FEF$_{50}$/FIF$_{50}$, a measure of extrathoracic airway obstruction, ranged from 12% to 67%, and the specificity ranged from 29% to 86%. The presence of the "sawtooth" sign on the flow volume loop , indicative of "pharyngeal" fluttering during forced breathing maneuvers, had a sensitivity ranging from 29% to 61%, and a specificity ranging from 54% to 85%. Using both FEF$_{50}$/FIF$_{50}$ and the "sawtooth sign" combined, the sensitivity ranged from 7% to 86%, and specificity from 13% to 89%. The meta-analysis of sensitivity and specificity yielded pooled estimates and ROC curve, as shown in Figure 7. The sensitivity of FEF$_{50}$/FIF$_{50}$ was 19.6% (± 9.6) and the specificity was 79.2% (±9.7). For the sawtooth sign, the sensitivity was 61.9% (±10.7) and the specificity 62.7% (±7.2). When both measures were analyzed together, the sensitivity was 39.1% (± 25.3) and specificity 60.5% (± 23.7).

Global Impressions: There were four studies (19A, 45A, 46A, 55A) reporting the global impression of clinicians: three studies in clinic settings (19A, 45A, 46A), and one (55A) in sleeping patients in a sleep laboratory. The evidence scores of these studies ranged from 19 to 28, and averaged 23.3. Together these studies included 1,139 patients total, 539 of whom were diagnosed with SA. AHI was the PSG metric used in three studies, with diagnostic cut-offs of 10 and 15. In the study of sleeping patients, the PSG metric was AI, and the cut-off for SA diagnosis was five. Their average age was 47.7 (three studies reporting), and the percentage of patients who were male was 67% (four studies reporting). The average BMI was 29.4 kg/m^2. Sensitivity of global impressions of SA relative to PSG diagnosis of SA ranged from 52% to 79%, with a pooled estimate of 58.9% (±4.2): specificity ranged from 50% to 100%, the latter result from the observation of sleeping patients. The pooled estimate of specificity was 65.6% (± 4.8). The summary ROC curve for these four studies is in Figure 8. These show that while sensitivity was relatively constant across studies, specificity varied a great deal.

Other Clinical: Nine studies (47A, 56A-63A) reporting sensitivity and specificity were identified for several clinical measures, but there were too few in each category to permit meta-analysis: neck circumference, airway dimensions via acoustic reflection (56A), nasopharyngeal airway resistance (57A), pulmonary function tests without flow volume curves (58A), laryngoscopy (59A), snoring sound analysis (60A), pupillary light reflex (61A), heart rate variability by ECG monitoring (62A), and body mass index alone (47A, 63A). No conclusions regarding the usefulness of any of these clinical measures as aids in the screening or diagnosis of SA can be made on the basis of so few studies.

Chemical: Similarly, there was one study (64A) of a chemical test (urinary uric acid and creatinine) as a screen for SA. There were 88 patients enrolled, 49 with SA. Patients who desaturated at night differed from those who did not, but no correlation can be made on the basis of a single study.

Radiologic: One study (65A) correlated MRI results with PSG. This study included 40 patients. One study (66A) correlated CT scans with PSG and included 37 patients. The latter also reported cephalometry, in relation to PSG results. There were additional cephalometry studies, which reported a multitude of different measurements of patients in different positions. Most of these studies did not, however, report correlations to PSG results, and none reported sensitivity or specificity in relation to PSG results. The average evidence score of these studies was 18.5 (range 17–20). Of the five radiologic studies which did report correlations to PSG results, one combined (49A) cephalometric results with morphometric results in a statistical model, which is discussed in the Prediction Equations section above. Among the remaining four studies (63A, 66A, 67A, 68A), 256 patients in total, there is too little overlap of measurements to pool data, or even to synthesize data in a strictly qualitative way.

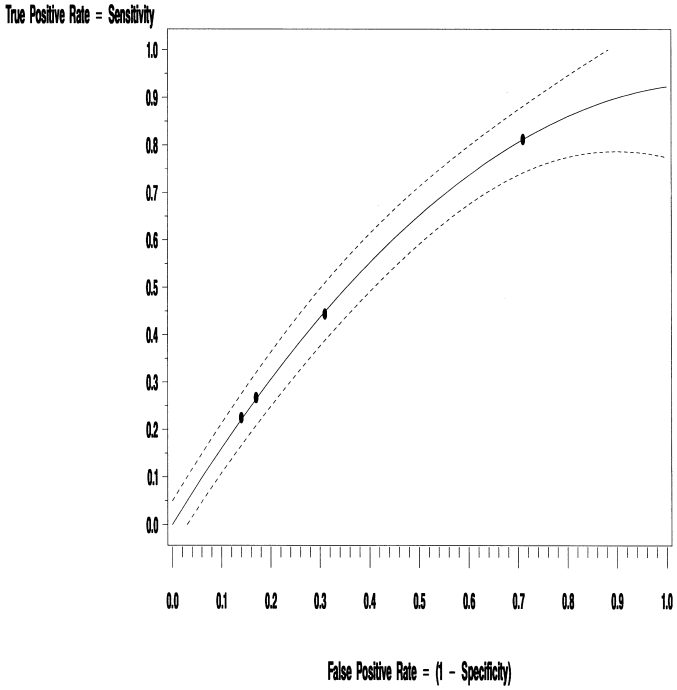
Questionnaires: Three studies (69A–71A) reported sensitivity, specificity, or correlations of focused questionnaires to PSG results. One of two studies (71A) which used the Epworth Sleepiness Scale (ESS), reported sensitivity (42%) and specificity (68%) in 354 suspected SA patients, using a PSG AI threshold of 20. The other ESS paper (69A) reported a correlation (r=0.55) to PSG RDI. The third paper (70A) did not use standard questionnaires, but selected questions about observed apneas, falling asleep or daytime sleepiness, and snoring. One additional paper[20] should be noted in this category, since it studied the sensitivity and specificity of several questions in a large number of patients (n=1,409). However, it only reported these outcomes by patient subgroups stratified by gender and as such it was not considered analyzable with the other studies in this set.

DISCUSSION

An ideal diagnostic test in a general population should have a relatively high specificity to minimize false positives, yet it should have sufficient sensitivity, and also be minimally intrusive, relatively inexpensive, possess "uniqueness" (i.e., not merely reflect other markers that are simpler or cheaper to acquire) and identify patients early in the disease process. Conversely, an ideal diagnostic test in a population with a high pre-test probability of disease should have higher sensitivity, while maintaining high specificity. Most of the patients in these studies were suspected of sleep apnea, and therefore had a higher pre-test probability of disease than the general population.

Differences among the studies in this set in sensitivity or specificity of identical tests vs. PSG may be due to several factors. A lower cutoff to declare a test positive in some studies may result in lower sensitivity or a higher cutoff may produce higher specificity. There may be random variations in the performance of the test between study sites, or between studies resulting in heterogeneity. There may be differences in the clinical settings in which the test is employed, or wide variability in the patient characteris-

Figure 7: Flow Volume Loops: FEF 50/FIF 50
Summary ROC Curve and 95% CI from Meta-Analysis of Individual Studies



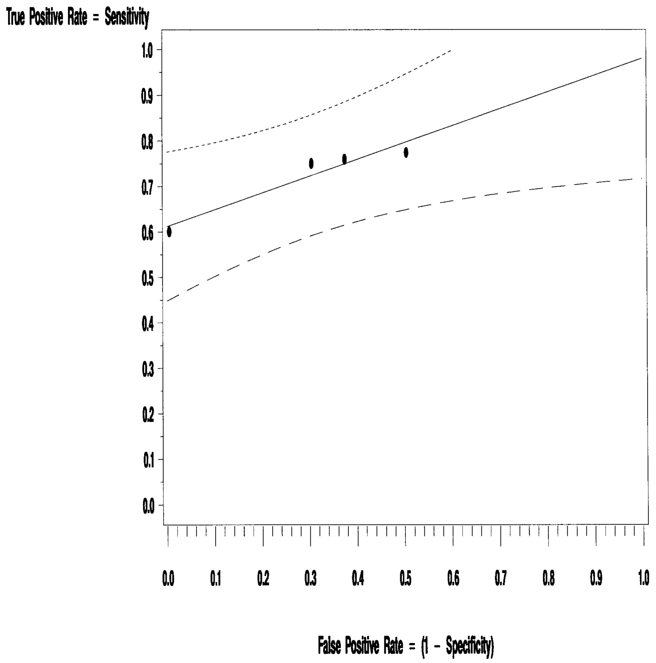
tics of those tested. Lastly, it should be recognized that the studies were frequently not designed to demonstrate comparability with a full PSG, but rather, to provide additional risk information or limited screening data.

The summary ROC curve serves as a compact description of the accuracy of a diagnostic test over a range of diagnostic thresholds.[16] These can also be used to compare technologies, to detect (and explore) outliers, and to build decision models. This is the first time summary ROC curves have been constructed for studies with sufficient data in SA. These curves indicate the degree of heterogeneity between the study sets, as well as the relationship between sensitivity and specificity within each study set. There is no ideal summary ROC curve for all clinical circumstances. The predictive value of any test depends upon the pre-test probability of the condition. Different centers with different referral populations will have different pretest probability, as will, of course, a general population for screening. Thus, these summary ROC curves should not necessarily dictate choice of tests. Also, in the absence of consideration of the pre-test probability, the costs of the test must be weighed in light of the size of the test population, and the costs and efficacy of treating cases thus identified.

In addition to the summary ROC curves, another key strength of this review is its comprehensiveness. This study represents the best available evidence base for the diagnosis of sleep apnea derived from relevant literature in five languages; and it is unlikely that any important diagnostic studies were missed. Restricting data analysis to those studies with features most likely to yield useful diagnostic test information, specifically to those studies reporting sensitivity and specificity of diagnostic maneuvers, is consistent with study selection approaches previously employed by the AASM Standards of Practice Committee in its 1994 statements,[12] in its 1997 statements[21,22] and the Blue Cross/Blue Shield Technology Evaluation Committee 1996 assessment of portable sleep studies. Furthermore, this established database, if kept updated, could serve as a valuable resource to practitioners and researchers.

Figure 8: Global Impression

Summary ROC Curve and 95% CI from Meta-Analysis of Individual Studies



There are many limitations to this evidence base which should be instructive to researchers planning and reporting new studies. In general, the diversity of designs and study objectives was high and the methodological rigor of the studies as an assessment of a diagnostic test was so low that, contrary to the usual practice of using evidence scores only in sensitivity analyses, these scores were used as a filter for selecting consistent studies for data extraction. Investigators thereby rejected studies scoring in the bottom 20% of the distribution of scores. Even so, the studies that remained constitute Level III to IV[23] evidence, that is, primarily derived from case series and observational studies. There were very few diagnostic studies which employed randomized assignment of tests, and very few studies performed blinded assessments of test results, both key features of rigorous diagnostic studies.

Numerous other limitations also apply to this dataset. With regard to the gold standard PSG, there was considerable variability in how PSGs were administered (i.e., which measures were considered essential components of standard PSGs). As a consequence, several questions are raised: Is the 'standard' PSG really a gold standard for the diagnosis of SA? Does the ability to measure sleep stage improve diagnostic accuracy? Is an entire night necessary? Proof is lacking, and the reasonably high sensitivity and specificity of partial channel PSGs and partial time PSGs only serve to reinforce this uncertainty. There was considerable inconsistency in how apnea and hypopnea were defined, let alone what metric (AI or AHI) and what threshold (>5, 10, 15, 20, 30 per hour) was used to diagnose SA. There was inconsistency in the incorporation of clinical signs and symptoms with PSG results in diagnosing SA. Distinctions between types of SA were usually not made. Night to night reproducibility of the gold standard is still not well documented, and may also differ using different diagnostic thresholds. In addition, usually AI and AHI were reported for the entire sleep duration, but these indices were virtually never reported separately for REM sleep and non-REM sleep. Although never formally studied, it remains possible that specific techniques have different sensitivities in detecting respiratory events during REM compared to during non-REM sleep (or indeed that respiratory events and arousals from REM sleep have different effects on the severity of symptoms than respiratory events and arousals from non-REM sleep.

Few studies included non-apneic patients, to achieve a broad spectrum of test subjects. Reliability of the PSG or diagnostic test being studied was not examined nor reported in these studies. Researchers should seek to clarify the prevalence of apnea and hypopnea in general populations by gender and age.[2] The Sleep Heart Health Study currently underway[24] will add to our knowledge of the prevalence of sleep apnea in a prospective cohort of 6,600 adults who will undergo a home PSG and be monitored for major cardiovascular events. More naturalistic sleep studies (in the home) are still of interest, as it is possible that much of the uncertainty about the nature of SA, its pathophysiology, risk factors, and clinical consequences, derives from the fact that the phenomenon called SA may be altered by the very fact of observing it via standard PSG.

Diagnostic technology is rapidly evolving, with increasing sensitivity of instruments and quite recent realization that very subtle respiratory events (e.g., changes in upper airway resistance and respiratory efforts causing recurrent arousals) can cause clinical symptoms, thus the frequency of diagnosis is changing.[25] Given these circumstances, and the aforementioned diversity in study design, methodology, and objectives, the use of common formats becomes essential. We recommend sleep apnea researchers adopt common reporting formats, in the spirit of that promoted for reporting randomized controlled trials.[26] Until such time as standardized diagnostic criteria for sleep-related breathing disorders are agreed upon among the professional societies and researchers in this field, as well as the insurance agencies, we recommend the following standardized report formats for researchers publishing results of diagnostic comparisons between standard PSG and other techniques in sleep apnea:

· Gold standard PSG should be performed in all patients over a full night
· Apnea and hypopnea criteria should be defined clearly and if various criteria are used, then the impact of varying these definitions on sensitivity and specificity should be noted
· AHI should be reported for total sleep time
· When reporting sensitivity and specificity, at least the standardized diagnostic AHI thresholds of <5, 5–30, and >30 should be used
· Patient groups should be defined using AHI alone vs AHI and clinical features
· The order of tests in diagnostic studies should be random
· Sleep monitoring systems proposed as pre-qualifiers or replacements to PSG must be validated in the settings in which they are intended to be used
· Test readers should be blinded to results of the other test
· The frequency of signs and symptoms (obesity, snoring, daytime sleepiness, observed apneas) should be noted
· The subject population should include a wide range of pre-test likelihood patients, including normals, and the prevalence in the study population should be assessed (i.e., pre-test likelihood of diagnosis)

With such standardization it will be possible to better combine future studies for useful comparison of techniques such that standardized criteria for diagnosis of sleep apnea can be developed and applied. Adoption of these recommendations, plus acceptance of research principles recently outlined by a Task Force of the American Academy of Sleep Medicine on defining sleep-related breathing disorders and measurement techniques[27] should not be delayed.

CONCLUSIONS

This systematic review of the best available evidence for diagnosis of sleep apnea suggests that although numerous diagnostic strategies have been reported, the published evidence for most is still insufficient as a basis for recommendations or guidelines. The following conclusions must be tempered by a recognition that reliance upon a full laboratory PSG as a gold standard is based upon a widely held, but unproven assumption. There is some evidence in a relatively small number of patients, that should be expanded with more studies, suggesting that a full laboratory PSG may not be necessary to diagnose SA. Rather, sleep laboratory measured oximetry, thoracoabdominal respiratory movements, and airflow alone (partial channel PSG), in the context of high likelihood of sleep apnea based upon features, may have sufficient sensitivity and specificity to replace full PSG. There is still insufficient evidence that any multi-channel portable device can be used reliably in the home setting. With regard to all the other types of assessments which were hoped to be somehow predictive of SA, including anthropomorphic signs, otolaryngeal and dental assessments, and radiologic measures, the etiologic relevance of most is at best controversial. While statistically significant associations have been noted in some studies, causal associations have not been proven. Flow volume loops do not appear to be a useful diagnostic test in SA. Lastly, sensitivity and specificity were good for clinical prediction rules in general, but additional studies would be required to build the evidence base to justify widespread adoption of any single model.

Future studies of diagnostic strategies should address the many limitations of the literature and in particular adopt standard research methods and reporting formats as offered herein.

ALTERNATE SOURCE

An executive summary of this report was posted on the Agency for Healthcare Policy and Research (AHCPR) Web site: (http://www.ahrq.gov/clinic/apnea.htm) in December 1998 and a hard copy of the report was published in

February 1999 (AHCPR Publication No. 99.E002) and posted on AHRQ's Web site: (http://text.nlm.nih.gov/ftrs/dbaccess/apnea). The report was reviewed in ACP J Club, 2000;132:69.

ACKNOWLEDGMENTS

We acknowledge the assistance of Harry Handelsman, DO, Task Order Officer, and Jacqueline Besteman, JD, MA, EPC Project Officer, in the Center for Practice and Technology Assessment, AHRQ. We acknowledge the assistance and dedication of our technical expert panel (TEP): Cynthia Barnes, RN, Ateo, NJ; Laurence Barsh, DMD, West Palm Beach, FL; Sudhansu Chokroverty, MD and James Stevens, MD, American Academy of Neurology, St Paul, MN; W. Ward Flemons, MD, University of Calgary and Foothills Hospital, Calgary, Alberta; David Hargett, A.W.A.K.E. Group, Wake-Up America, and The A.P.N.E.A. Network, Bolingbrook, IL; Jeffrey Lipsitz, MD, Sleep Disorders Centre of Metropolitan Toronto, Toronto, Ontario; Anne Meneghetti, MD, Blue Cross Blue Shield of Massachusetts, Boston, MA; Allan Pack, MB, CHB, PhD, Center for Sleep and Respiratory Neurobiology, Hospital of the University of Pennsylvania, Philadelphia, PA; Aaron Sher, Capital Region Otolaryngology Head and Neck Group, Albany, NY; James Stevens, Fort Wayne Neurological Center, Fort Wayne, IN; J. Woodrow Weiss, MD, Beth Israel-Deaconess-East, Sleep Disorders Center, Boston, MA; and Terry Young, PhD, University of Wisconsin-Madison, Department of Preventive Medicine, Madison, WI.

FINANCIAL DISCLOSURE:

This article was developed under contract 290-97-0016 with the agency for Healthcare Research and Quality (AHRQ) formerly the Agency for Health Care Policy and Research (AHCPR).

The information reflects solely the opinion(s) of the author(s), and nothing should be contrived to indicate AHCPR's support or endorsement of its content.

REFERENCES

1. Phillipson EA. Sleep apnea — a major public health problem (Editorial). N Engl J Med 1993;328:1271-3.

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

3. Teran-Santos J, Jimenez-Gomes, A, Cordero-Guevara J, The Cooperative Group Burgos-Santander. The association between sleep apnea and the risk of traffic accidents. N Engl J Med 1999;340:847-51.

4. National Commission on Sleep Disorders Research. Wake up America: A national sleep alert. Washington, D.C.: U.S. Government Printing Office, 1993.

5. He J, Kryger MH, Zorick FJ, Conway W, Roth T. Mortality and apnea index in obstructive sleep apnea. Experience in 385 male patients. Chest 1988;94:9-14.

6. Pack AI. Simplifying the diagnosis of obstructive sleep apnea (editorial; comment). Ann Intern Med 1993;119:528-9.

7. Hosselet JJ, Norman RG, Ayappa I, Rapoport DM. Detection of flow limitation with a nasal cannula/pressure transducer system. Am J Respir Crit Care Med 1998;157:1461-67.

8. 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.

9.Mulrow CD, Oxman AD eds. Cochrane collaboration handbook (updated 9 December 1996). The cochrane collaboration; issue 1. Oxford: Update Software, 1997.

10. Mulrow C, Cook DJ, Davidoff F. Systematic reviews: critical links in the great chain of evidence. Ann Intern Med 1997;156:376-80.

11. Sacks HS, Berrier J, Reitman D, Ancona-Berk V, Chalmers T. Meta-analyses of randomized controlled trials. N Engl J Med 1987;316:450-55

12. Ferber R, Millman R, Coppola M, Fleetham J, Murray C, Iber C, McCall V, Nino-Murcia G, Pressman M, Sanders M, Strohl K, Votteri B, Williams A. Portable recording in the assessment of obstructive sleep apnea. Sleep 1994;17 (4):378-92.

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

14. Flemons WW, Remmers JE. Methods of Diagnosing Sleep Apnea — the diagnosis of sleep apnea: questionnaires and home studies. Sleep 1996;19(10):S243-S7.

15. Fleiss, J. Statistical methods for rates and proportions. New York:John Wiley & Sons, 1973.

16. Littenberg B, Moses LE. Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method. Med Decision Making 1993;13:313-21.

17. Littenberg B, Mushlin AI, and the Diagnostic Technology assessment consortium. Technetium bone scanning in the diagnosis of Osteomyelitis: A meta-analysis of test performance. J Gen Intern Med 1992;7:158-163.

18. Moses LE, Shapiro D. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993;12:1293-1316.

19. Rechtschaffen A and Kales A eds. A manual of standardised terminology, techniques and scoring system for sleep stages of human subjects. Los Angeles, CA: UCLA, BIS/BRI, 1968.

20. 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.

21. Chesson AL, Ferber RA, Fry JM, Grigg-Damberger M, Hartse KM, Hurwitz TD, Johnson S, Kader GA, Littner M, Rosen G, Sangal B et al. An American Sleep Disorders Association Report. Practice parameters for the indications for polysomnography and related procedures. Sleep 1997;20(6):406-22.

22. Chesson AL, Ferber RA, Fry JM, Grigg-Damberger M, Hartse KM, Hurwitz TD, Johnson S, Kader GA, Littner M, Rosen G, Sangal B et al. An American academy of sleep medicine review: the Indications for polysomnography and related procedures. Sleep 1997;20(6):423-87.

23. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest 1992;102:305S-11S.

24. Quan SF, Howard BV, Iber C, et al. The sleep heart health study: design, rationale, and methods. Sleep 1997;20(12):1077-1085.

25. Guilleminault C, Stoohs R, Clerk A, Simmons J, and Labanowski M. From obstructive sleep apnea syndrome to upper airway resistance syndrome: Consistency of daytime sleepiness. Sleep 1992; 15:S13-6.

26. Begg C, Cho M, Eastwood S, Horton R, et al. Improving the quality of reporting of randomized controlled trials-The CONSORT Statement. JAMA 1996; 276: 637-9.

27. 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;22:667-89.

APPENDIX A

1A. Carmona Bernal C, Capote Gil F, Cano Gomez S, Sanchez Armengol A, Medina Gallardo JF, Castillo Gomez J. Brief polysomnographic studies in the diagnosis of the obstructive sleep apnea syndrome. Arch. Bronconeumol 1994;30:390-93.

2A. Fanfulla F, Patruno V, Bruschi C, and Rampulla C. Obstructive sleep apnoea syndrome: is the "half-night polysomnography" an adequate method for evaluating sleep profile and respiratory events? Eur Respir J 1997;10:1725-1729.

3A. Persson HE and Svanborg E. Sleep deprivation worsens obstructive sleep apnea. Comparison between diurnal and nocturnal polysomnography. Chest 1996;109:645-650.

4A. Sanders MH, Black J, Costantino JP, Kern N, Studnicki K, Coates J. Diagnosis of sleep-disordered breathing by half-night polysomnography. Am Rev Respir Dis 1991;144:1256-61.

5A. Series F, Cormier Y, and La Forge J. Validity of diurnal sleep recording in the diagnosis of sleep apnea syndrome. Am Rev Respir Dis 1991;143:947-9.

6A. Van Keimpema ARJ, Rutgers SR, Strijers RLM. The value of one hour daytime sleep recording in the diagnosis of sleep apnea syndrome. J Sleep Res 1993;2:257-59.

7A. Scharf SM, Garshick E, Brown R, Tishler PV, Tosteson T, McCarley R. Screening for subclinical sleep-disordered breathing. Sleep 1990;13:344-53.

8A. Garcia Diaz EM, Capote Gil F, Cano Gomez S, Sanchez Armengol A, Carmona Bernal C, Soto Campos JG. Respiratory polygraphy in the diagnosis of obstructive sleep apnea syndrome. Arch Bronconeumol 1997;33:69-73.

9A. Carrasco O, Montserrat JM, Lloberes P, Ascasco C, Ballester E, Fornas C, Rodriguez-Rosin R. Visual and different automatic scoring profiles of respiratory variables in the diagnosis of sleep apnoea-hypopnoea syndrome. Eur Respir J 1996;9:125-30.

10A. Lloberes P, Montserrat JM, Ascaso A, Parra O, Granados A, Alonso P, Vilaseca I, Rodriguez-Roisin R. Comparison of partially attended night time respiratory recordings and full polysomnography in patients with suspected sleep apnoea/hypopnoea syndrome. Thorax 1996;51:1043-47.

11A. Douglas NJ, Thomas S, Jan MA. Clinical value of polysomnography. Lancet 1992;339:347-350.

12A. Duchna HW, Rasche K, Orth M, Schultze-Werninghaus G. Sensitivity and specificity of pulse oximetry in diagnosis of sleep-related respiratory disorders. Pneumologie 1995;49 (Suppl 1):113-5.

13A. Farney RJ, Walker LE, Jensen RL, Walker JM. Ear oximetry to detect apnea and differentiate rapid eye movement (REM) and non-REM (NREM) sleep. Screening for the sleep apnea syndrome. Chest 1986;89:533-9.

14A. Levy P, Pepin JL, Deschaux-Blanc C, Paramelle B, Brambilla C. Accuracy of oximetry for detection of respiratory disturbances in sleep apnea syndrome. Chest 1996;109:395-9.

15A. Rodriguez Gonzalez-Moro JM, de Lucas Ramos P, Sanchez Juanes MJ, Izquierdo Alonso JL, Peraita Adrados R, Cubillo Marcos JM. Usefulness of the visual analysis of night oximetry as a screening method in patients with suspected clinical obstructive sleep apnea syndrome. Arch Bronconeumol 1996;32:437-41.

16A. Series F, Marc I, Cormier Y, La Forge J. Utility of nocturnal home oximetry for case finding in patients with suspected sleep apnea hypopnea syndrome. Ann Intern Med 1993;119:449-53.

17A. Yamashiro Y and Kryger MH. Nocturnal oximetry: Is it a screening tool for sleep disorders? Sleep 1995;18:167-71.

18A. Gugger M. Comparison of resmed autoset (version 3.03) with polysomnography in the diagnosis of the sleep apnoea/hypopnoea syndrome. Eur Respir J 1997;10:587-91.

19A. Gyulay S, Olson LG, Hensley MJ, King MT, Allen KM, Saunders NA. A comparison of clinical assessment and home oximetry in the diagnosis of obstructive sleep apnea. Am Rev Respir Dis 1993; 147:50-3.

20A. Svanborg E, Larsson H, Carlsson-Nordlander B, Pirskanen R. A limited diagnostic investigation for obstructive sleep apnea syndrome. Oximetry and static charge sensitive bed. Chest 1990;98:1341-5.

21A. Pepin JL, Levy P, Lepaulle B, Brambilla C, Guilleminault C. Does oximetry contribute to the detection of apneic events? Mathematical processing of the SaO_2 signal. Chest 1991;99:1151-7.

22A. Acebo C, Watson RK, Bakos L, Thoman EB. Sleep and apnea in the elderly: reliability and validity of 24-hour recordings in the home. Sleep 1991;14:56-64.

23A. Bradley PA, Mortimore IL, Douglas NJ. Comparison of polysomnography with rescare autoset in the diagnosis of the sleep apnoea/hypopnoea syndrome. Thorax 1995;50:1201-3.

24A. Emsellem HA, Corson WA, Rappaport BA, Hackett S, Smith LG, Hausfeld JN. Verification of sleep apnea using a portable sleep apnea screening device. South Med J 1990;83:748-752.

25A. Esnaola S, Duran J, Infante-Rivard C, Rubio R, Fernandez A. Diagnostic accuracy of a portable recording device (MESAM IV) in suspected obstructive sleep apnoea. Eur Respir J 1996;9:2597-2605.

26A. Finke R, Jurczok A, and Matthys H. Clinical experience with the apnea check system in screening for sleep apnea. Pneumologie 1993;47 (Suppl 1):119-21.

27A. Fleury B, Rakotonanahary D, Hausser-Hauw C, Lebeau B, Guilleminault C. A laboratory validation study of the diagnostic mode of the Autoset system for sleep-related respiratory disorders. Sleep 1996; 19:502-5.

28A. Gyulay S, Gould D, Sawyer B, Pond D, Mant A, Saunders N. Evaluation of a microprocessor-based portable home monitoring system to measure breathing during sleep. Sleep 1987;10:130-42.

29A. Hamm M, Krause J, Felsmann M, Barnstorf D, Kothe R, Fabel H. A computerized processing unit for ambulatory diagnosis of sleep apnea and nocturnal hypoxemia. Pneumologie 1990;44 (Suppl 1):627-8.

30A. Hida W, Shindoh C, Miki H, Kikuchi Y, Okabe Shinchi, Taguchi O, Takishima T, Shirato K. Prevalence of sleep apnea among Japanese industrial workers determined by a portable sleep monitoring system. Respiration 1993;60:332-7.

31A. Issa FG, Morrison D, Hadjuk E, Iyer A, Feroah T, Remmers JE. Digital monitoring of sleep-disordered breathing using snoring sound and arterial oxygen saturation. Am Rev Respir Dis 1993; 148:1023-9.

32A. Kiely JL, Delahunty C, Matthews S, McNicholas WT. Comparison of a limited computerized diagnostic system (ResCare AUTOSET) with polysomnography in the diagnosis of obstructive sleep apnoea syndrome. Eur Respir J 1996;9:2360-2364.

33A. Koziej M, Cieslicki JK, Gorzelak K, Sliwinski P, Zielinski J. Hand-scoring of MESAM 4 recordings is more accurate than automatic analysis in screening for obstructive sleep apnoea. Eur Respir J 1994; 7:1771-1775.

34A. Man GC, and Kang BV. Validation of a portable sleep apnea monitoring device. Chest 1995;108:388-393.

35A. Parra O, Garcia-Esclasans N, Montserrat JM, Eroles LG, Ruiz J, Lopez JA, Guerra JM, Sopena JJ Should patients with sleep apnoea/hypopnoea syndrome be diagnosed and managed on the basis of home sleep studies? Eur Respir J 1997; 10:1720-4.

36A. Rauscher H, Popp W, Zwick H. Quantification of sleep disordered breathing by computerized analysis of oximetry, heart rate and snoring. Eur Respir J 1991;4:655-9.

37A. Redline S, Tosteson T, Boucher MA, Millman RP. Measurement of sleep-related breathing disturbances in epidemiologic studies. Assessment of the validity and reproducibility of a portable monitoring device. Chest 1991;100:1281-6.

38A. Roos M, Althaus W, Rhiel C, Penzel T, Peter JH, von Wichert P. Comparative use of MESAM IV and polysomnography in sleep-related respiratory disorders. Pneumologie 1993;47 (Suppl 1):112-118.

39A. Stoohs R, Guilleminault C. MESAM 4: an ambulatory device for the detection of patients at risk for obstructive sleep apnea syndrome (OSAS). Chest 1992;101:1221-7.

40A. Stoohs R, Guilleminault C. Investigations of an automatic screening device (MESAM) for obstructive sleep apnoea. Eur Respir J 1990; 3:823-9.

41A. Tvinnereim M, Mateika S, Cole P, Haight J, Hoffstein V. Diagnosis of obstructive sleep apnea using a portable transducer catheter. Am J Respir Crit Care Med 1995;152:775-9.

42A. White DP, Gibb TJ, Wall JM, Westbrook PR. Assessment of accuracy and analysis time of a novel device to monitor sleep and breathing in the home. Sleep 1995;18:115-26.

43A. Zucconi M, Ferini-Strambi L, Castronovo V, Oldani A, Smirne S. An unattended device for sleep-related breathing disorders: validation study in suspected obstructive sleep apnoea syndrome. Eur Respir J 1996;9:1251-6.

44A. 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;64:194-99.

45A. Viner S, Szalai JP, Hoffstein V. Are history and physical examination a good screening test for sleep apnea? Ann Intern Med 1991; 115:356-59.

46A. Hoffstein V, Szalai JP. Predictive value of clinical features in diagnosing obstructive sleep apnea. Sleep 1993;16:118-22.

47A. Vaidya AM, Petruzzelli GJ, Walker RP, McGee D, Gopalsami C. Identifying obstructive sleep apnea in patients presenting for laser-assisted uvulopalatoplasty. Laryngoscope 1996;106:431-37.

48A. Kushida CA, Efron B, Guilleminault C. A predictive morphometric model for the obstructive sleep apnea syndrome. Ann Intern Med 1997;127:581-87.

49A. Pracharktam N, Nelson S, Hans MG, Broadbent BH, Redline S, Rosenberg C, Strohl KP. Cephalometric assessment in obstructive sleep apnea. Am J Orthod Dentofacial Orthop 1996;109:410-19.

50A. Quera-Salva MA, Guilleminault C, Partinen M, Jamieson A. Determinants of respiratory disturbance and oxygen saturation drop indices in obstructive sleep apnoea syndrome. Eur Respir J 1988; 1:626-31.

51A. Hoffstein V, Wright S, Zamel N. Flow-volume curves in snoring patients with and without obstructive sleep apnea. Am Rev Respir Dis 1989;139:957-60.

52A. Krieger J, Weitzenblum E, Vandevenne A, Stierle JL, Kurtz D. Flow-volume curve abnormalities and obstructive sleep apnea syndrome. Chest 1985;87:163-67.

53A. Rauscher H, Popp W, Zwick H. Flow-volume curves in obstructive sleep apnea and snoring. Lung 1990;168:209-14.

54A. Shore ET Millman RP. Abnormalities in flow-volume loop in obstructive sleep apnea sitting and supine. Thorax 1984;39:775-9.

55A. Haponik EF, Smith PL, Meyers DA, Bleecker ER. Evaluation of sleep-disordered breathing. Is polysomnography necessary? Am J Med 1984;77:671-77.

56A. Katz I, Stradling J, Slutsky AS, Zamel N, Hoffstein V. Do patients with obstructive sleep apnea have thick necks? Am Rev Respir Dis 1990; 141:1228-31.

57A. Suratt PM, McTier RF, Wilhoit SC. Collapsibility of the nasopharyngeal airway in obstructive sleep apnea. Am Rev Respir Dis 1985; 132:967-71.

58A. Onal E, Leech JA, and Lopata M. Relationship between pulmonary function and sleep-induced respiratory abnormalities. Chest 1985;87:437-41.

59A. Geibel M, Schonhofer B, Rolzhauser HP, Wenzel M, Kohler D. Predictive value of laryngoscopy with reference to the severity of obstructive sleep apnea. Pneumologie 1997;51 (Suppl 3):809-10.

60A. Fiz JA, Abad J, Jane R, Riera M, Mananas MA, Caminal P, Rodenstein D, Morera J. Acoustic analysis of snoring sound in patients with simple snoring and obstructive sleep apnoea. Eur Respir J 1996; 9:2365-70.

61A. Pressman MR, Fry JM. Relationship of autonomic nervous system activity to daytime sleepiness and prior sleep. Sleep 1989;12:239-45.

62A. Keyl C, Lemberger P, Pfeifer M, Hochmuth K, Geisler P. Heart rate variability in patients with daytime sleepiness suspected of having sleep apnoea syndrome: a receiver-operating characteristic analysis. Clin Sci (Colch) 1997;92:335-43.

63A. Lowe AA, Fleetham JA, Adachi S, Ryan CF. Cephalometric and computed tomographic predictors of obstructive sleep apnea severity. Am J Orthod Dentofacial Orthop 1995;107:589-95.

64A. Braghiroli A, Sacco C, Erbetta M, Ruga V, Donner CF. Overnight urinary uric acid: creatinine ratio for detection of sleep hypoxemia. Validation study in chronic obstructive pulmonary disease and obstructive sleep apnea before and after treatment with nasal continuous positive airway pressure. Am Rev Respir Dis 1993;148:173-78.

65A. Rodenstein DO, Dooms G, Thomas Y, Liistro, G, Stanescu DC, Culee C, Aubert-Tulkens G. Pharyngeal shape and dimensions in healthy subjects, snorers, and patients with obstructive sleep apnoea. Thorax 1990;45:722-27.

66A. Shinohara E, Kihara S, Yamashita S, Yamane M, Nishida M, Arai T, Kotani K, Nakamura T, Takemura K, Matsuzawa Y. Visceral fat accumulation as an important risk factor for obstructive sleep apnoea syndrome in obese subjects. J Intern Med 1997;241:11-18.

67A. Davies RJ, Stradling JR. The relationship between neck circumference, radiographic pharyngeal anatomy, and the obstructive sleep apnoea syndrome. Eur Respir J 1990;3:509-14.

68A. Will MJ, Ester MS, Ramirez SG, Tiner BD, McAnear JT, Epstein L. Comparison of cephalometric analysis with ethnicity in obstructive sleep apnea syndrome. Sleep 1995;18:873-5.

69A. Johns MW. A new method for measuring daytime sleepiness:The epworth sleepiness scale. Sleep 1991;14:540-5.

70A. 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.

71A. Pouliot Z, Peters M, Neufeld H, Kryger MH. Using self-reported questionnaire data to prioritize OSA patients for polysomnography. Sleep 1997;20:232-6.