Sleep, 20(11):991–1001

Sleep-Disordered Breathing

Automated Detection and Classification of Sleep-Disordered Breathing From Conventional Polysomnography Data

*B. H. Taha, *J. A. Dempsey, †S. M. Weber, †‡M. S. Badr, †‡J. B. Skatrud, *T. B. Young, *A. J. Jacques and *K. C. Seow

**Department of Preventive Medicine, University of Wisconsin–Madison;*
†Department of Medicine, University of Wisconsin–Madison; and
‡Medical Service, William S. Middleton Memorial Veterans Hospital, Madison, Wisconsin, U.S.A.

Summary: Efficient automated detection of sleep-disordered breathing (SDB) from routine polysomnography (PSG) data is made difficult by the availability of only indirect measurements of breathing. The approach we used to overcome this limitation was to incorporate pulse oximetry into the definitions of apnea and hypopnea. In our algorithm, 1) we begin with the detection of desaturation as a fall in oxyhemoglobin saturation level of 2% or greater once a rate of descent greater than 0.1% per second (but less than 4% per second) has been achieved and then ask if an apnea or hypopnea was responsible; 2) an apnea is detected if there is a period of no breathing, as indicated by sum respiratory inductive plethysmography (RIP), lasting at least 10 seconds and coincident with the desaturation event; and 3) if there is breathing, a hypopnea is defined as a minimum of three breaths showing at least 20% reduction in sum RIP magnitude from the immediately preceding breath followed by a return to at least 90% of that "baseline" breath. Our evaluation of this algorithm using 10 PSG records containing 1,938 SDB events showed strong event-by-event agreement with manual scoring by an experienced polysomnographer. On the basis of manually verified computer desaturations, detection sensitivity and specificity percentages were, respectively, 73.6 and 90.8% for apneas and 84.1 and 86.1% for hypopneas. Overall, 93.1% of the manually detected events were detected by the algorithm. We have designed an efficient algorithm for detecting and classifying SDB events that emulates manual scoring with high accuracy. **Key Words:** Sleep-disordered breathing—Polysomnography—Computerized detection.

Polysomnography (PSG) is the current standard method for the assessment of sleep-disordered breathing (SDB). Manual scoring of apneas and hypopneas from an overnight PSG study can require several hours of tedious event identification and tabulation. Furthermore, this time-consuming analysis is naturally subject to errors and inconsistencies relating to human subjectivity and fatigue, as demonstrated by the interscorer variability reported in the literature (1,2).

Efforts to overcome some of the problems associated with manual scoring have led to the introduction of computerized and computer-assisted SDB scoring systems. The last few years have seen the advent of many algorithms and computerized devices for the automated assessment of sleep (3). To date, there have been few extensive studies to validate the algorithms of these systems. Furthermore, most of the available validation studies are correlative and do not involve an event-by-event validation that matches actual computer-detected events with manually detected ones.

Although these advances have contributed greatly to the assessment and management of patients with SDB, they have left assessment of SDB in nonclinical research settings, where sleep apnea is less prevalent and hypopneas dominate, in need of comprehensive automated detection and reporting of SDB events.

To improve the efficiency and accuracy of sleep assessment using conventional PSG, we developed a computer algorithm for detecting and classifying apneas and hypopneas. We reasoned that we needed to begin with one of the more reliable signals available

Accepted for publication August 1997.

Address correspondence and reprint requests to Jerome A. Dempsey, Ph.D., Department of Preventive Medicine, University of Wisconsin–Madison, 504 North Walnut Street, Madison, WI 53705, U.S.A.

TABLE 1. *Subject characteristics and apnea-hypopnea index (AHI) values*

Age (year)	Height (cm)	Weight (kg)	Sex	AHI
53	173	85	M	2
52	160	91	F	14
40	168	68	M	10
57	185	103	M	19
40	178	84	M	40
55	198	125	M	7
56	178	107	F	51
48	198	113	M	26
62	147	107	F	9
55	174	66	F	4

M, male; F, female.

The AHI is based on the routine manual scoring definition of apnea (at least 10 seconds of no airflow) and hypopnea (50% reduction in airflow associated with at least 4% desaturation).

to us, namely oxyhemoglobin saturation, and then assess the accuracy of computerized desaturation detection and determine if a sleep-disordered event was associated with it. The algorithm uses desaturations of 2% or more as event markers of both apneas and hypopneas. It then analyzes the breathing pattern to detect, classify, and tabulate the causing event. We evaluated the performance of the algorithm by comparing its results to manual scoring on an event-by-event basis.

METHODS

Subjects

PSG records from 10 participants in the Wisconsin sleep cohort study—a community-based longitudinal study of the natural history of SDB (4)—were selected to equally represent low [apnea–hypopnea index (AHI) 0–10 per hour], medium (AHI 10–20), and high (AHI greater than 20) SDB severity levels. The technical quality of the signals in these selected studies was typical of those obtained in the entire cohort. Table 1 shows the subject characteristics and the apnea–hypopnea index (AHI = number of apneas and hypopneas per hour of sleep) from conventional manual scoring.

Data acquisition

Conventional PSG consisted of continuous polygraphic recording from surface electrodes for central and occipital electroencephalography (EEG), right and left electrooculography (EOG), chin and leg electromyography (EMG), and electrocardiography (ECG), and from noninvasive sensors for oro-nasal airflow (thermistors), nasal airflow (using an infrared CO_2 detector), tracheal sounds (microphone), respiratory inductance plethysmography (RIP), and oxyhemoglobin saturation level by pulse oximetry (Spo_2) using a finger probe. The transducers and lead wires permitted normal positional changes during sleep. Bedtime and awakening time were at each subject's discretion. All signals were low-pass filtered and sampled at 64 Hz on-line using an IBM-compatible 386DX-25MHz computer equipped with a data acquisition board (LabMaster DMA, Scientific Solutions, Solon, MA). Sampled data were then transferred to CD-ROM for permanent storage. Figure 1 shows some of the above signals in a sleep record from a subject exhibiting SDB.

The RIP signal was calibrated by first instructing the subject to perform an isovolume maneuver while adjusting the relative gains of the abdomen and rib cage component signals such that a net zero sum RIP was obtained. Following this adjustment the subject was instructed to breathe through a spirometer at increasing tidal volumes. The sum RIP signal was then calibrated against the spirometer readings using a linear regression equation (5).

PSG paper records were manually scored for sleep and movement in 30-second periods. Sleep data were staged [stages I, II, III, IV, and rapid eye movement (REM) sleep] according to the system of Rechtschaffen and Kales (6). Only periods of sleep were analyzed for apnea and hypopnea detection.

Operational definitions of detected events

Breath detection

A breath was defined from the digitized sum RIP signal as the period from the start of one inspiration to the start of the next inspiration. The detection was performed by smoothing the sum RIP signal (three-sample average) and then differentiating it to obtain a pseudoflow signal. A flow rate tolerance of 25 ml per second was applied to the baseline to determine the zero flow crossings corresponding to the start and end of inspiration and expiration. For each detected breath, the following parameters were computed: inspiratory time (T_I), expiratory time (T_E), breathing frequency per minute (f_b), tidal volume (V_T = average of inspiratory and expiratory volumes), inspiratory duty cycle [$T_I/(T_I + T_E)$], and minute ventilation ($\dot{V}_E = V_T \times f_b$). Breath detection was similarly performed on the rib cage and abdominal RIP signals.

Desaturation

Figure 2 shows a schematic diagram summarizing the rules for detecting a desaturation. The algorithm analyzes the digitized Spo_2 signal, averaged every 0.5

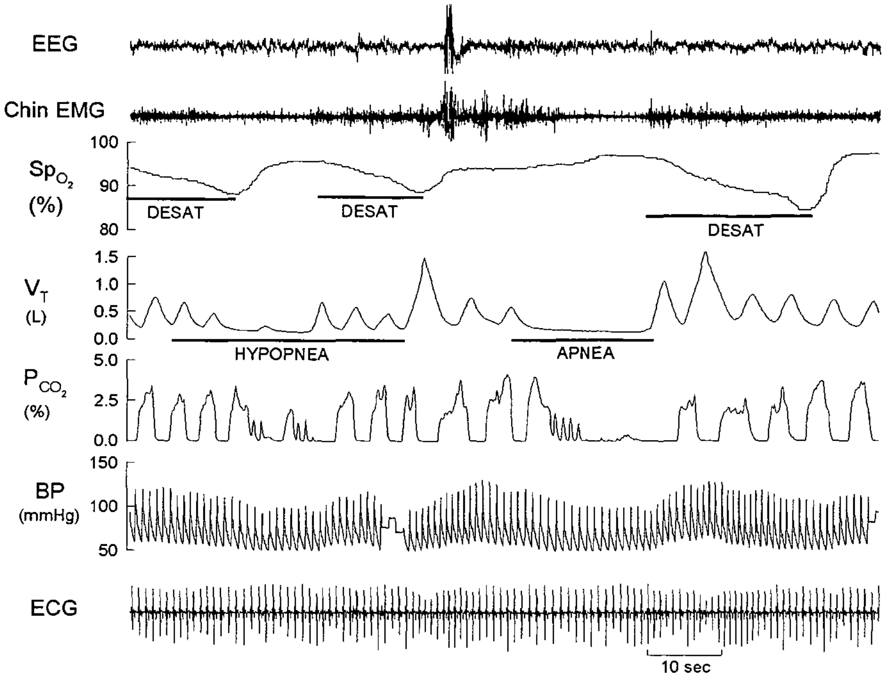

FIG. 1. Examples of computer-detected apnea and hypopnea. Note the delay in oxygen saturation and the increased blood pressure and heart rate in response to the sleep-disordered breathing events. EEG, electroencephalogram; EMG, electromyogram; Spo_2, oxyhemoglobin saturation determined by pulse oximetry; V_T, tidal volume from sum RIP; Pco_2, expired CO_2 concentration; BP, blood pressure; ECG, electrocardiogram.

second, and detects a desaturation when the levels marked *a*, *b*, and *c* in Fig. 2 are identified. These levels are defined to avoid false detection of artifactual fluctuations in the Spo_2 signal. Level *a* is the point at which Spo_2 achieves a rate of fall greater than 0.1% per second and less than 4% per second; at *b*, Spo_2 achieves a minimum at least 2% below *a*; and at *c*, Spo_2 returns to a level either 1% below *a* or 3% above *b*, whichever occurs sooner. The total time from *a* to *c* must be between 10 and 60 seconds.

Apnea

An apnea is defined simply as a period of no inspiration as indicated by the differentiated sum RIP signal (see above) lasting 10–60 seconds.

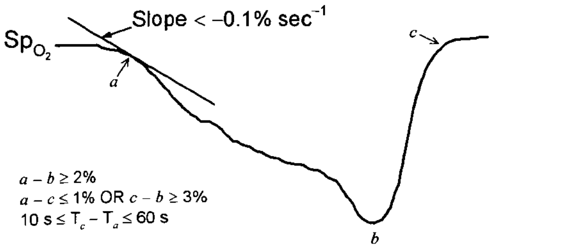
Slope < -0.1% sec^{-1}
Sp_{O_2}
a
c
b
$a - b \geq 2\%$
$a - c \leq 1\%$ OR $c - b \geq 3\%$
$10\ s \leq T_c - T_a \leq 60\ s$

FIG. 2. Schematic diagram illustrating the rules for detecting a desaturation. Point *a* indicates the point at which a rate of fall greater than 0.1% per second is first achieved, *b* is the minimum signal level attained, and *c* is a level that is 3% or more above *b* or 1% or less below *a*. T*a* and T*c* are the times of levels *a* and *c*, respectively.

Hypopnea

Figure 3 illustrates the rules for detecting a hypopnea. A hypopnea is identified by the algorithm when the following sequence of events is found: 1) a breath has a magnitude less than 80% of the immediately preceding breath (marked B in Fig. 3); 2) the next two breaths at least also have magnitudes below 80% of B; and 3) a breath has a magnitude at least 90% of B and starts less than 180 seconds after the start of breath B.

We used event-by-event comparison of computer-detected and manually detected hypopneas to arrive at

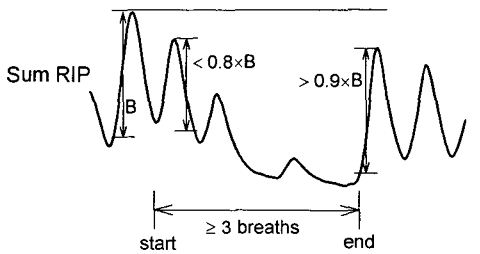

FIG. 3. Schematic diagram illustrating the computer definition of a hypopnea. Vertical arrows indicate tidal volumes computed as the averages of inspiratory and expiratory volumes of each detected breath. B indicates the tidal volume of the "baseline" breath. A hypopnea starts with a breath having a sum respiratory inductance plethysmography (RIP) magnitude below 80% of the previous breath (B), followed by at least two more consecutive breaths also having RIP magnitudes below 80%. The hypopnea is then terminated with the first breath achieving a RIP magnitude above 90% of B, provided it occurs within 3 minutes.

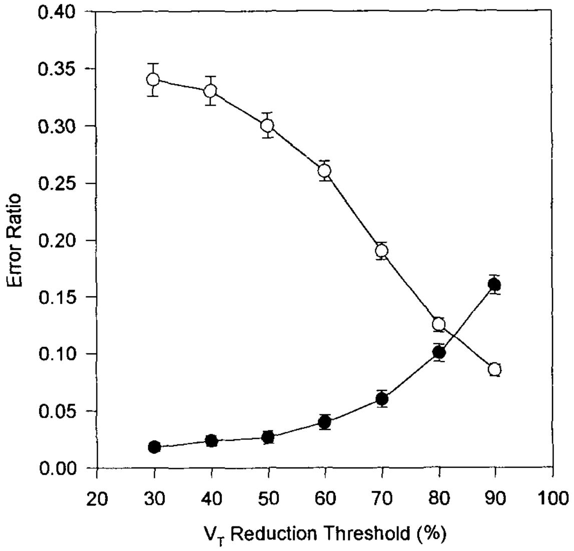

FIG. 4. Inclusion and exclusion errors at different tidal volume (V_T) reduction thresholds for hypopnea detection. Inclusion error is the length of the period during a computer-detected hypopnea that is not part of the manually detected hypopnea, expressed as a fraction of the computer-detected hypopnea duration. Exclusion error is the length of the period during the manually detected event that is not part of the computer-detected event, expressed as a fraction of the manually detected event duration. The filled circles are inclusion errors, and the open circles are exclusion errors. The point of intersection of the two curves offers the best performance compromise.

the 80% V_T reduction rule for hypopnea definition. The analysis consisted of computing two types of errors per computer event. Inclusion error is the length of the period during a computer-detected hypopnea that is not part of the manually detected hypopnea, expressed as a fraction of the computer-detected hypopnea duration. Conversely, exclusion error is the length of the period during a manually detected event that is not part of the computer-detected event, expressed as a fraction of the manually detected event duration. An event that is totally missed by the computer was assigned the maximum inclusion and exclusion error value of 1.0, and an event that was exactly matched was assigned zero errors. Next, V_T reduction threshold was varied from 30 to 90%, and each time, the average inclusion and exclusion errors were computed for all events in all subjects. The resultant averaged errors are shown in Fig. 4. As the reduction threshold became greater, inclusion error increased and exclusion error decreased. We chose the V_T reduction threshold closest to the point of intersection of the two error curves (80%) as the point representing the best compromise. Figure 4 shows that computer-based hypopneas at 80% V_T reduction have average inclusion and exclusion errors of 12%.

Event classification

Apneas were classified into central, mixed, or obstructive by determining if there were any abdominal and rib cage breathing efforts during the apnea. If there were compartmental breathing movements throughout the apnea duration, then the apnea was classified as obstructive because the rib cage and abdominal motion would have had to be in complete paradox to result in a zero sum signal. If there were no detected rib cage or abdominal breaths for at least one-fourth of the apnea duration, the apnea was classified as mixed. Finally, an apnea with no thoracic or abdominal breathing efforts throughout its duration was classified as central.

Manual event detection

For the purpose of validating our algorithm, both the analog paper and digital PSG records were manually scored. An expert polysomnographer scored the 10 records according to the following protocol (see Fig. 1): 1) The manual scorer first examined the computer-detected desaturation events (see above) to determine if, in the scorer's judgment, they were actual physiological desaturation events or artifactual. 2) Once a physiological desaturation event was confirmed, apnea was defined as a period of no breathing lasting at least 10 seconds that was contiguous and consistent with the desaturation event's timing. The signals used to detect airflow were the oro-nasal thermistor, the nasal end-tidal CO_2 monitor, and the sum RIP signals. Each apnea was classified by the polysomnographer as central, mixed, or obstructive according to the same criteria employed by the computer algorithm (see above). 3) If a desaturation was detected and no apnea was detected, the sum RIP signal was examined. A hypopnea was identified when any discernible decrease in sum RIP magnitude occurred contiguously with the observed desaturation. The amount of reduction judged sufficient was left to the scorer's experience to determine if the pattern of change in the RIP amplitude was physiological and relevant to the desaturation.

The manual scorer had access to both digital and analog versions of all the recorded PSG signals during the scoring process and was able to manipulate the signal gain and temporal resolution of the digitized signals. The manual scoring was done prior to computerized apnea and hypopnea detection for all records. Manually detected events were entered into 30-second epoch-segmented computer files in a format synchronized with the digitized signals to facilitate event-by-event comparison.

Computer-based event detection

The operation of the computer algorithm followed the same logic as the manual procedure. For every

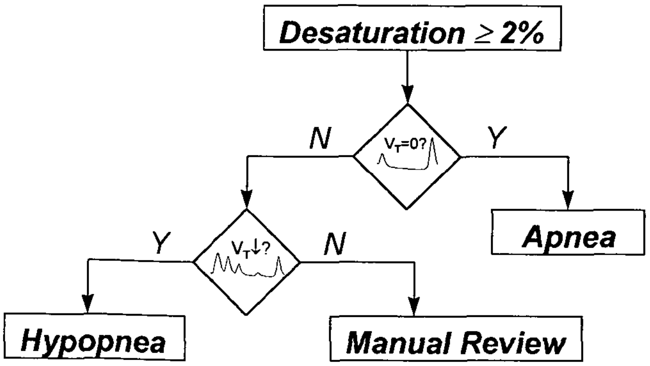

FIG. 5. Flow chart of detection algorithm. Signals in the decision boxes are typical representations of respiratory inductance plethysmography (RIP) apnea and hypopnea patterns. Y, yes; N, no.

detected desaturation event, the program checked for the presence of an apnea first and a hypopnea second to account for desaturation. If neither was found, the desaturation event was flagged for subsequent manual review. Figure 5 shows a high-level flow chart of the algorithm's operation. The following rules were applied during the procedure: 1) Due to the variable delay between the onset of an SDB event and the consequent desaturation (7), apneas and hypopneas were restricted to occur within a window beginning 40 seconds prior to the start of the desaturation and ending at the time of minimum saturation. This range spanned all delays from manually scored SDB events to their associated desaturations. 2) An apnea was allowed to start after the start of a desaturation event because it is possible (and common) to have a reduction in V_T before complete cessation of breathing takes place. A hypopnea, however, had to start before the start of desaturation. 3) If multiple apneas were found within the search period, then the one closest to the desaturation start time was registered as the responsible event. If multiple hypopneas were found, the one with a start time closest to 16 seconds earlier than the desaturation start time was chosen. This time interval was chosen because it was the mean delay time from the beginning of a manually scored hypopnea to its associated desaturation.

Validation

We evaluated the performance of the detection algorithm in three phases. First, we determined the desaturation detection performance by computing the positive predictive value (PV+). This value indicates the probability that a computer-detected event was a true event and is given by

$$PV+ = \frac{TP}{TP + FP} \times 100$$

where TP (true positives) is the number of computer-detected desaturation events confirmed by the manual scorer and FP (false positives) is the number of events determined by the manual scorer to be erroneously detected by the computer.

The second phase of evaluation dealt with the detection accuracy of apneas and hypopneas. We applied sensitivity and specificity analyses to apnea detection, hypopnea detection, and overall apnea plus hypopnea detection. In addition to the parameters defined for desaturation above (TP and FP), we defined FN (false negatives) as manually scored breathing events that were missed by the computer and TN (true negatives) as cases of confirmed desaturation events not associated with SDB events by either manual or computer-based scoring. This approach was justifiable because SDB events were considered only if desaturation events were detected. This definition permitted the calculation of sensitivity and specificity as

$$Se = \frac{TP}{TP + FN} \times 100$$

$$Sp = \frac{TN}{TN + FP} \times 100$$

and negative predictive value as

$$PV- = \frac{TN}{TN + FN} \times 100.$$

Finally, we evaluated the algorithm's performance for apnea classification by comparing the manual and algorithm results.

RESULTS

Desaturation detection

There were 1,938 true positive (TP) and 58 false positive (FP) desaturation events, yielding a positive predictive value of 97.1%. Sixteen of the desaturations identified as false positive were due to body movement artifacts in which the artifact rejection provisions of the algorithm failed. Movement was easily identifiable by the manual scorer as myogenic and electromechanical artifact in other channels coincident with the desaturation event. In the remaining 42 FP desaturation events, the algorithm registered a fall in Spo_2 of less than 2%. The manual scorer was able to attribute everyone of the TP desaturation events to either an apnea (258 events) or a hypopnea (1,680 events).

The following sections describe the degree of agreement between computerized and manual scoring of apneas and hypopneas and explain the sources of disagreement.

TABLE 2. *Parameters of algorithm performance vs. manual scoring*

	TP	FN	FP	TN				
	Number of events				% Se	% Sp	% PV+	% PV−
Apnea	190	68	161	1588	73.6	90.8	54.1	95.9
Hypopnea	1412	268	78	484	84.1	86.1	94.8	64.4
SDB events[a]	1804	134	58	—	93.1	—	96.9	—

SDB, sleep-disordered breathing; TP, true positive; FN, false negative; FP, false positive; TN, true negative; Se, sensitivity; Sp, specificity; PV+, positive predictive value; and PV−, negative predictive value.

[a] TN for SDB analysis refers to the section of the sleep record that contained no SDB events (not a countable event).

Apnea detection

The first row of Table 2 shows the results of the event-by-event comparison of computerized versus manual scoring of apneas. The following detection parameters are shown: TP, apneas detected manually and by the algorithm; FN, apneas detected manually but not by the algorithm (including apneas detected as hypopneas by the algorithm); FP, apneas detected by the algorithm but not manually (including manually detected hypopneas), and TN, desaturation events that were not attributed to apneas by both the algorithm and the manual scorer.

The 68 FN apneic events included 42 computer-detected hypopneas and 26 events for which the algorithm did not identify either an apnea or a hypopnea. One of the 161 FP events was due to an FP desaturation, and the remaining 160 were manually detected hypopneas (rather than apneas). We discuss the reasons for these disagreements below.

Hypopnea detection

The second row of Table 2 shows the same parameters for hypopnea detection as those described for apnea. The 268 FN hypopneic events included 160 computer-detected apneas and 108 events missed by the algorithm. Thirty-six of the 78 FP hypopneic events were due to FP desaturation events, and the remaining 42 were manually scored apneas. Most of the reasons for apnea detection disagreements, discussed below, also explain the disagreements here.

Overall SDB detection

The last row of Table 2 shows the results of the comparison between manual and computerized scoring of SDB events, regardless of whether they were apneas or hypopneas. TP now refers to manually scored SDB events also detected by the computer (including apneas detected as hypopneas and vice versa). FN events are manually scored SDB events missed by the algorithm.

TABLE 3. *Summary of reasons for disagreement between manually detected and computer-detected SDB events*

Error	Reason for disagreement	% Total number of errors
Missed apnea	No inspiration for <10 seconds	7.7
Missed hypopnea	$<80\%$ reduction in V_T over several breaths	25.9
	Movement artifact	6.3
Apnea detected as hypopnea	Algorithm detected breaths, manual scorer did not	3.3
	No breathing for less than 10 seconds, but breathing pattern satisfies hypopnea detection criteria	9.2
Hypopnea detected as apnea	Manual scorer detected breaths, algorithm did not	32.1
	Algorithm detected sum RIP clipping artifact as apnea	15.5

SDB, sleep-disordered breathing; V_T, tidal volume; RIP, respiratory inductance plethysmography.

Total number of disagreement errors was 336.

All 58 FP SDB events are due to FP desaturation events. This is because the manual scorer was able to attribute every TP desaturation to an SDB event, and thus there were no desaturation events for which the algorithm detected an SDB event and the manual scorer did not. In the context of this analysis, TN refers to periods during which neither the manual scorer nor the computer found any SDB events, which is a noncountable feature by definition because it constitutes the segments of the sleep record containing no SDB events.

Reasons for detection disagreement

Completely missed events

Table 3 shows a breakdown of the reasons for events missed by the algorithm. All 26 apneas completely missed by the algorithm had periods of no inspiration that were slightly shorter than 10 seconds and were therefore rejected by the algorithm's criteria for apnea duration. It is important to note, however, that these events were of sufficient duration to cause a detectable desaturation that was verified manually (8.3–9.7 seconds long).

Of the 108 hypopneas that the algorithm completely missed, 87 were manually scored hypopneas in which the reduction in RIP amplitude occurred slowly over several breaths, none of which was quite 20% less than the previous one, thus not meeting the algorithm detection requirements. The other 21 events included movement artifacts in the RIP signal that violated the algorithm's criteria for hypopnea detection but that did

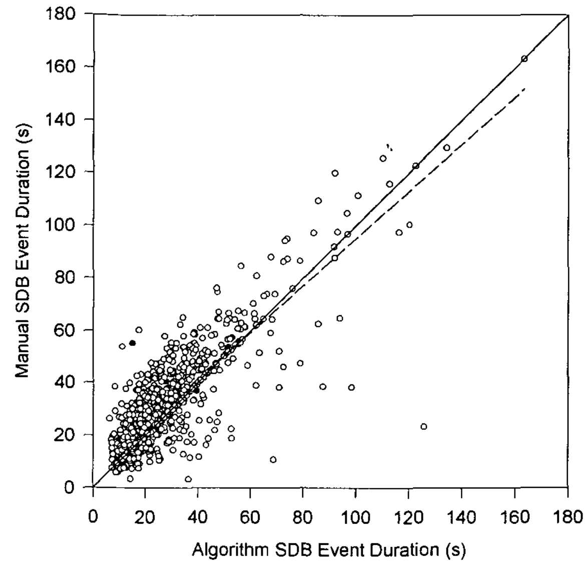

FIG. 6. Relationship of computer-detected sleep-disordered breathing (SDB) event duration to manually detected events. Open circles are manually scored hypopneas, and filled circles are manually scored apneas. The solid line is the line of identity, and the dashed line is the linear regression line.

not prevent the manual scorer from still identifying the breathing reduction pattern.

All of these missed events (108 + 26 = 134) were flagged by the algorithm for manual review by the algorithm as "unexplained" desaturation events (see Methods).

Misclassified SDB events

The FP and FN apnea and hypopnea entries in Table 2 include apneas detected as hypopneas and vice versa. As Table 3 shows, this misclassification was mostly due to disagreement between the computerized and the manual scoring on whether breathing was present (in 35.4% of the erroneously detected SDB events). In these cases, small changes in the sum RIP signal were interpreted differently by the manual scorer and the algorithm, where one determined that there was cessation of breathing and the other detected breathing.

Two other situations occurred that contributed to the misclassification error. In 9.2% of the erroneously detected SDB events, there was a period of breathing cessation (as seen on the sum RIP signal) that occurred within a hypopneic pattern and that was slightly shorter than 10 seconds. These were scored manually as apneas, whereas the algorithm ruled the apnea out (because of the <10-second duration) and consequently detected a hypopnea. Finally, 15.5% of the SDB events in error were due to the appearance of a signal-clipping artifact (due to digitization) that did not prevent the manual scorer from observing a hypopnea but caused the algorithm to falsely detect an apnea.

TABLE 4. *Computer classification of true positive apneas*

	Computer		
Manual	Obstructive	Central	Mixed
Obstructive	97	0	45
Central	2	31	8
Mixed	0	0	7

Manual versus computer event durations

Figure 6 shows a scatter plot of the manual versus TP computer-detected SDB event durations. There is reasonable agreement between the two events (r = 0.85). However, manual events were slightly but significantly longer than computer-detected events (28.5 + 17 vs. 25.4 + 16.3 seconds, $p < 0.05$).

Apnea classification

Of the 190 TP apneas, 142 were obstructive, 41 central, and 7 mixed, as determined by manual scoring. As shown in Table 4, the computer classified these apneas into 99 obstructive, 31 central and 60 mixed events. The percentage of apneas correctly classified was 71.1%. There was an overall tendency of the algorithm to overestimate mixed apneas. This was due to more specific computer rules for determining mixed apneas compared with the largely subjective determination by visual inspection of the signals.

DISCUSSION

Algorithm design considerations

We made several algorithm design decisions based on our understanding of SDB and guided by our SDB event characterization requirements. In this section, we discuss the rationale for our design decisions as well as the underlying limitations.

Use of desaturation

Our decision to use desaturation as the initial event in the detection and classification of apneas and hypopneas was guided by several factors. One primary reason stems from our confidence in the desaturation measurement and detection techniques we used. As our validation indicates, desaturation events can be detected with high positive predictability. We believe that the excellent performance of the algorithm in detecting physiological desaturation is due to the measures taken to improve Spo_2 signal quality and artifact rejection. First, the pulse oximeter was set to produce an average value every 3 seconds (fast mode) to maximize desaturation level tracking in time. Second, the digitized signal (at 64 Hz) was averaged every 0.5 second to improve signal-to-noise ratio and provide data reduction. Finally, we added several checks into the algorithm for artifact rejection, including maximum rate of fall and minimum duration requirements.

The second justification for using desaturation information lies in its physiological importance. Given the qualitative or, at best, semiquantitative nature of commonly used measurements of breathing (thermistors, expired CO_2 monitors, and respiratory inductive plethysmography), most sleep centers further qualify the reduction of airflow (or breathing movement) by requiring a certain degree of desaturation (8). Further justification for this approach can be found in numerous studies that report good correlation between desaturation and apnea–hypopnea frequencies (9–14). This good correlation was found even when assessing desaturation without any amplitude criteria dealing only with saturation cyclical changes (15). The manual scorer in this study was able to attribute all detected physiological desaturation events to apneas and hypopneas. Similarly, the algorithm was able to attribute 93.1% of the desaturation events to SDB events.

The use of desaturation also has practical benefits. The feature of interest in the saturation signal (desaturation event) is relatively easy to extract because of its minimal features. It simply consists of a reduction and a subsequent rise in level. This ease of recognition has obvious implications for artifact rejection. Extracting features responsible for the same respiratory event from another, more complex signal, such as RIP, would involve more complicated rules and result in a higher false-positive rate. In fact that was precisely our experience in early attempts to define hypopnea only on the basis of V_T reduction patterns in the sum RIP signal.

Use of the 2% level, as opposed to commonly used greater levels of desaturations, to qualify apneas and hypopneas was prompted by our desire not to overlook any potentially significant SDB events. This decision was made in light of our confidence that the detected desaturations, down to the 2% level, were all "real". Douglas et al. (16) reported that some patients with documented mild to moderate sleep apnea can have events associated with desaturations not exceeding 2%, with some even benefiting from continuous positive airway pressure (CPAP) treatment. Furthermore, there is evidence that episodes of high upper airway resistance may not be associated with any measurable desaturation but lead to an arousal from sleep (17). George et al. (10) used computer detection of desaturations of 3% or more to accurately predict the number of apneas and hypopneas. In addition, our intended application of the algorithm to a study of a working population made it essential to detect less severe breathing events than would be required in a clinical setting. Note that the absolute value measurement error of ±2% reported for the pulse oximeter (18) does not impact our algorithm, which detects relative changes in Spo_2 level and uses strict criteria for acceptance.

Although the successful use of desaturation alone as a predictor of the apnea-hypopnea index may be helpful in the clinical diagnosis of SDB (10,12,13), it is not an adequate index for many research studies. It offers no information about the SDB event itself, such as duration or type (i.e. apnea or hypopnea, central or obstructive apnea). This information is essential to the understanding of the causes and effects of these events.

Apnea and hypopnea definitions

Apnea definition is usually based on measurements of airflow, namely the oro-nasal thermistor or thermocouple output and the nasal CO_2 monitor. However, both of these signals offer only qualitative indications of airflow. The thermistor and thermocouple simply sense changes in temperature near the mouth and nose in response to inspiratory and expiratory airflow. Temperature changes can also occur for reasons unrelated to breathing, such as head movement. The self-adhesive band holding the thermistor in place often slips from the mouth and nose during prolonged sleep studies. In a recent study, Whyte et al. (2) demonstrated that measurements of oro-nasal airflow may have little added value in the presence of reliable RIP signals. Similarly, the CO_2 monitor tubing often becomes obstructed by condensed expired water vapor and ceases to provide a useful signal.

The RIP signals (rib cage, abdomen, and sum), on the other hand, are robust indices of respiratory effort. Although they do not directly measure airflow, reduced or no-airflow episodes are almost always recognizable in a well-calibrated and properly processed RIP sum signal. By definition, central apnea is associated with zero effort and would therefore produce a flat (zero) sum RIP signal. In addition, an episode of upper airway obstruction sufficient to produce apnea is associated with phase-reversed thoracoabdominal motion, and addition of the two RIP components cancels out to yield a flat (zero) sum RIP. Another advantage in using the RIP signals to indicate apneas is the ability to classify them as central or obstructive according to the presence of any compartmental breathing motion (see the Methods section). Most important, we found the use of the RIP signal to be essential to achieving acceptable sensitivity for apnea detection. We compared RIP signal usefulness to that of the CO_2 and thermistor signals for apnea detection. The apnea detection sensitivity of 73.6% with the RIP signal reported here (see Table 2) was reduced to 41% when the CO_2 and thermistor signals were used. Other investigators have reported the use of the sum RIP signal to detect apneas (2,19).

There are also problems with the RIP signal that limit its use. It has been shown that the calibration of the RIP signal against a spirometer is not usually maintained throughout the night and is drastically affected by postural changes (20). In addition, the rib cage and abdomen respiratory bands are subject to movement and slippage throughout the night, especially in obese subjects. Furthermore, as with all signals, the RIP is subject to movement artifacts and baseline drift. For these reasons, we did not attempt to use the RIP signal as an indicator of absolute tidal volume at any point in our algorithm. As described above, the algorithm analyzes the signal with attention only to relative magnitude changes to recognize hypopneas.

Objective investigation of the amount of relative RIP magnitude reduction required to declare a hypopnea has been addressed by Gould et al. (14). In their study, the numbers of hypopneas detected per hour using several levels of sum RIP magnitude reduction (25, 50, and 75%) were compared with the number of EEG arousals per hour and the number of 4% desaturations per hour. Their analysis led them to choose 50% reduction in V_T to indicate a hypopnea. Although such an analysis provides an objective assessment of hypopneas, it suffers, in our view, from a major drawback. It is a correlative method based on the number of events per hour for each subject with no provision to ensure event-to-event correspondence. That is, there is no assurance that different physiological events are being compared, although their totals per hour may be similar. We used event-by-event comparison of computer-detected and manually scored hypopneas to determine the optimal V_T reduction for hypopnea definition (see the Methods section). This procedure for determining the operating parameters of our algorithm overcomes the limitation of correlative methods. However, it remains for us to determine whether our hypopneas are indeed physiologically significant (see below).

Finally, it should be noted that the use of the RIP analysis criteria described above to detect SDB events without reference to desaturation results in the extreme overestimation of their frequency (especially hypopneas). However, such criteria are needed to achieve the high overall sensitivity reported here (93.1%) once the desatuaration events have been detected. This approach to using the RIP signal (as a secondary qualification for SDB detection) overcomes the many limitations inherent in this measurement.

Design limitations and implications

Requirement of desaturation for SDB detection

An obvious possibility resulting from the desaturation requirement for SDB detection is to miss a true SDB that does not produce a desaturation. A low minimum desaturation, 2%, was chosen precisely to minimize this possibility. Given the many precautions we used to obtain a faithful and sensitive SpO_2 signal (fast-mode acquisition, high sampling rate, and artifact rejection), we believe that there will be few, if any, hypopneas of physiological significance that do not cause at least a 2% desaturation. In any case, for such events to be labeled physiologically significant there must be some physiological indications, such as arousals or cardiovascular responses, which currently either preclude automated detection or are not part of conventional PSG. The current algorithm uses only data available during conventional PSG studies and requires no manual intervention aside from routine sleep scoring. Within these parameters, it offers an attractive alternative to manual scoring in similar population studies, as evidenced by its good detection characteristics.

Use of digitized data in manual scoring

As described above in the Methods section, the manual scoring was performed on digitized data, and the manual scorer was able to manipulate signal gain and viewing length. This contrasts with the conventional method for manual scoring using paper records. We opted to use digital data because we wanted to test the algorithm's performance against the best human "judgment" concerning the occurrence of an SDB event. We did not want to put the manual scorer at a disadvantage caused by the limited signal range available on paper records. Requiring the manual scorer to make a judgment based on less data than the algorithm had access to would have resulted in an underestimation of the algorithm's ability to detect SDB events that would reflect not its performance but rather the less than optimal "gold standard" used.

Application to research studies

Research studies, usually those of populations, often have quite different aims than diagnostic clinical tests of patients with sleep-related complaints. Often the questions addressed in population studies are different from those pursued by systems developed for use in sleep clinics or for home monitoring. In a longitudinal population study involving thousands of sleep records, such as the Wisconsin sleep cohort study (4), the need for efficient and consistent SDB detection cannot be overemphasized. The detection algorithm developed and validated here is a good tool that we plan to use for this purpose. It offers consistency, objectivity, and efficiency not available with manual scoring.

Implications and future direction

Whether the hypopneas we detected are physiologically significant, and whether we missed any significant events are questions that deserve further investigation. However, it is important to note that the aim of this work was to automate a manual procedure for purposes of efficiency and consistency. We believe that we offer an algorithm that achieves very good agreement with manual scoring in addition to sensitive detection of mild to severe SDB events. Our overall detection sensitivity (93.1%) compares well with that reported by George et al. (10). In their study, which to our knowledge is the only one that has performed an event-by-event analysis, they reported that automatic detection of desaturations levels of 3% or more achieved a 97.9% sensitivity in detecting apneas and hypopneas. However, they used a much more sensitive device for saturation measurement (Hewlett-Packard ear oximeter) that is not generally used in PSG because of its large probe size and high cost and because it is no longer in production. In addition, their algorithm detected only desaturation events and provided no information on the SDB events that caused them or their classification.

The specific apnea and hypopnea detection sensitivities were reasonably good (73.6 and 84.1%, respectively). Most misclassification errors were due to subtle deviations from the specific detection criteria of the algorithm, such as the 10-second apnea requirement or the minimum sum RIP change needed to indicate a breath (Table 3). Such deviations were very difficult to detect visually by the manual scorer. It is feasible and indeed likely that the manual scorer, given the precise computer measurement of timing, would revise the decision to rule in some of the apneas because the periods of breathing cessation were indeed shorter than 10 s. Had we allowed such a revision in our analysis we would have increased the apnea detection sensitivity from 73.6 to 83.7%.

Another source of disagreement between the manual scoring and the computer scoring can be attributed to the fact that there were more breathing indicators available to the manual scorer than to the algorithm. This gave the manual scorer more contextual clues to the nature of the SDB events. With that in mind, it can be argued that the algorithm accuracy can be improved by incorporating more signals (such as thermistor and CO_2 monitor signals) into the definitions of apnea and hypopnea. However, it is not a trivial matter to incorporate multiple signals into a single definition, especially if that is to be done in a manner that emulates the manual scoring practice of examining several traces concurrently. We believe that the use of multiple signals to indicate SDB requires techniques that are more sophisticated than the simple sequential stringing of conditions. The principles of fuzzy logic have such a quality and may provide useful tools for the detection and analysis of SDB using multiple physiological indicators. It is worth noting, however, that one of the most attractive features of the algorithm presented in this study is its simplicity—the use of only two signals (Spo_2 and sum RIP) to detect apneas and hypopneas very reliably.

CONCLUSION

We have developed a computer algorithm for the automated detection and classification of apneas and hypopneas from conventional PSG data. We have validated the algorithm's performance on an event-by-event basis and showed it to be accurate and efficient. Further investigation into the physiological significance of computer-detected hypopneas is needed to assess the clinical and diagnostic usefulness of this method of detection.

Acknowledgements: This study was a project of the University of Wisconsin–Madison Sleep and Respiration Research Group. Additional coinvestigators in the University of Wisconsin Sleep Center of Research for sleep-disordered breathing include Mari Palta and Nizar Jarjour. We are also indebted to Leah Steinberg, Deborah Brown, Andrea Darner, Kathy Kennison, Hyon Kim, Kathryn Pluff, Linda Evans, Ron Leder, and Robert Mulroy for their technical assistance. The National Institutes of Health (PO1-HL 42242 and RPO 3186) and the Veterans Administration Medical Research Foundation funded the original research.

REFERENCES

1. Lord S, Sawyer B, Pond D, et al. Interrater reliability of computer-assisted scoring of breathing during sleep. *Sleep* 1989; 12(6):550–8.
2. Whyte KF, Allen MB, Fitzpatrick MF, Douglas NJ. Accuracy and significance of scoring hypopneas. *Sleep* 1992;15(3):257–60.
3. Zimmerman JT, Torch WC, Reichert JA. A comparison of sleep-data-acquisition-and-analysis systems and computerized-paperless polysomnography. *J Polysomnographic Technol* 1992:30–44.
4. 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.
5. Sackner MA, Watson H, Belsito AE, et al. Calibration of respiratory inductive plethysmography during natural breathing. *J Appl Physiol* 1989;66(1):410–20.
6. Rechtschaffen A, Kales A. *A manual of standardized terminology, techniques and scoring systems for sleep stages of human*

subjects. Bethesda, Maryland: U.S. Government Printing Office, 1968.

7. West P, George CF, Kryger MH. Dynamic in vivo response characteristics of three oximeters: Hewlett-Packard 47201A, Biox III, and Nellcor N-100. *Sleep* 1987;10(3):263–71.
8. Moser NJ, Phillips BA, Berry DT, Harbison L. What is hypopnea, anyway? *Chest* 1994;105(2):426–8.
9. 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(4):1023–9.
10. George CF, Millar TW, Kryger MH. Identification and quantification of apneas by computer-based analysis of oxygen saturation. *Am Rev Respir Dis* 1988;137(5):1238–40.
11. 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(6):655–9.
12. Pépin JL, Lévy P, Lepaulle B, Brambilla C, Guilleminault C. Does oximetry contribute to the detection of apneic events? Mathematical processing of the SaO2 signal. *Chest* 1991;99(5): 1151–7.
13. Rauscher H, Popp W, Zwick H. Computerized detection of respiratory events during sleep from rapid increases in oxyhemoglobin saturation. *Lung* 1991;169(6):335–42.
14. Gould GA, Whyte KF, Rhind GB, et al. The sleep hypopnea syndrome. *Am Rev Respir Dis* 1988;137:895–8.
15. Sériès F, Marc I, Cormier Y, La Forge J. Nocturnal home oximetry for the sleep apnea-hypopnea syndrome. *Ann Intern Med* 1993;119:449–53.
16. Douglas NJ, Thomas S, Jan MA. Clinical value of polysomnography. *Lancet* 1992;339(8789):347–50.
17. Guilleminault C, Stoohs R, Clerk A, Cetel M, Maistros P. A cause of excessive daytime sleepiness. The upper airway resistance syndrome. *Chest* 1993;104(3):781–7.
18. Ohmeda. *Ohmeda Biox 3740 Pulse Oximeter operating/maintenance manual*. Louisville, CO: Ohmeda, 1990.
19. Bradley TD, Martinez D, Rutherford R, et al. Physiological determinants of nocturnal arterial oxygenation in patients with obstructive sleep apnea. *J Appl Physiol* 1985;59(5):1364–8.
20. Whyte KF, Gugger M, Gould GA, Molloy J, Wraith PK, Douglas NJ. Accuracy of respiratory inductive plethysmograph in measuring tidal volume during sleep. *J Appl Physiol* 1991;71(5): 1866–71.