https://doi.org/10.5664/jcsm.8278

JCSM | Journal of Clinical Sleep Medicine

SCIENTIFIC INVESTIGATIONS

A comparison of automated and manual sleep staging and respiratory event recognition in a portable sleep diagnostic device with in-lab sleep study

Zhigang Zhang, MD[1,2]; Mudiaga Sowho, MD[1]; Tamas Otvos, MD[1]; Larissa Sanglard Sperandio, MD[1]; Joshua East, RPSGT[1]; Frank Sgambati[1]; Alan Schwartz, MD[1]; Hartmut Schneider, MD, PhD[1]

[1]Department of Medicine, Division of Pulmonary and Critical Care Medicine, Johns Hopkins University, Baltimore, Maryland; [2]Department of Geriatrics, Peking University First Hospital, Beijing, China

Study Objectives: The objectives were to develop and validate an algorithm for editing WatchPAT scoring and assess the accuracy in an unselected clinical population as well as age and sex substrata.

Methods: Two hundred sixty-two participants were enrolled to undergo WatchPAT simultaneously with in-lab polysomnography (PSG) recordings for developing (n = 30), optimizing (n = 62), and validating (n = 170) an algorithm to review and edit respiratory events and sleep architecture of WatchPAT recordings, which was based on visual inspection of WatchPAT signals. Apnea-hypopnea index (AHI) and sleep indices were compared with PSG-derived and automated WatchPAT indices.

Results: Although estimation of total sleep time (TST) was comparable between automated and manual algorithm, estimation of rapid eye movement (REM) sleep time was markedly improved with manual editing from 0.48, 23.0 min (−43.9 to 89.8) to 0.64, 18.3 min (−32.6 to 69.1) (correlation with PSG, mean difference [reference range] from PSG, respectively). Manual scoring also improved correlation and agreement with PSG AHI from 0.65, 2.5 events/h (−24.0 to 28.9) to 0.81, −4.5 events/h (−22.5 to 13.6) as well as concordance for categorical agreement of sleep-disordered breathing severity and concordance for detecting severe REM-related sleep-disordered breathing. Interscorer reliabilities were excellent for TST and AHI, while good for REM sleep time. The automated algorithm performed better in younger than in older patients, while performed similarly between men and women with respect to concordance statistics. The manual algorithm markedly improved concordances more in older patients and women than in their counterparts.

Conclusions: Our manual editing algorithm improves correlation and agreement with PSG-derived sleep and breathing indices. Sex and age influence the accuracy of automated analysis and the performance of manual editing on AHI concordance.

Keywords: home sleep apnea test, peripheral arterial tone, respiratory events, sleep-disordered breathing, sleep staging

Citation: Zhang Z, Sowho M, Otvos T, et al. A comparison of automated and manual sleep staging and respiratory event recognition in a portable sleep diagnostic device with in-lab sleep study. *J Clin Sleep Med*. 2020;16(4):563–573.

BRIEF SUMMARY

Current knowledge/Study Rationale: WatchPAT is an approved home sleep test device based on peripheral arterial tone. Yet methods for implementing guidelines for reviewing of WatchPAT recordings are not available. We sought to develop and validate an algorithm for manual editing WatchPAT scoring and to assess its accuracy in both an unselected clinical population and subpopulations of different ages and sex.

Study Impact: This study shows that manual editing is reliable and improves the agreement with PSG-derived indices and that sex as well as age influences both the accuracy of automated analysis and the performance of manual editing on AHI concordance. Its use may facilitate efficient diagnosis and severity estimation of all patients with sleep-disordered breathing.

INTRODUCTION

Sleep-disordered breathing (SDB) is a common, increasingly recognized medical condition worldwide and is characterized by repeated episodes of apnea or hypopnea during sleep, leading to daytime sleepiness and cognitive dysfunction and increased risk of hypertension, diabetes, stroke, and cancer.[1–4] An essential public health priority is the diagnosis and severity estimation of SDB. The gold standard for the diagnosis of SDB remains attended overnight polysomnography (PSG). Nevertheless, PSG has several disadvantages, including its relatively high cost, requirement for full in-laboratory or hospital testing, excess burden from multiple sensors, an unfamiliar sleep environment, and backlog into the laboratory. Therefore, practical constraints have placed greater emphasis on conducting home sleep apnea tests (HSAT) across medical entities.

Many HSAT devices have been developed as an alternative to full PSG for diagnosis of SDB. Most HSAT devices detect respiratory events by monitoring airflow, respiration effort, and oxygen saturation, but do not assess sleep/wake or sleep stages. WatchPAT (Itamar Ltd, Israel) has incorporated a unique set of algorithm to stage sleep and recognize SDB events from patient's oxygen saturation, sympathetic tone (peripheral arterial tone [PAT] and heart rate changes) and actigraphy. Several studies and meta-analyses[5–9] have demonstrated its

accuracy of diagnosing SDB, leading to a recent approval for its use as a HSAT device by the American Academy of Sleep Medicine (AASM). Nonetheless, general adoption of WatchPAT in clinical routine has been hindered by the use of unconventional signals (no flow, respiratory effort, or electroencephalography), making it difficult for physicians and respiratory technologists to review, verify, and interpret WatchPAT recordings. Moreover, the aforementioned validation studies did not reflect the spectrum of patients presenting to clinical sleep centers, thus raising questions about its generalizability and accuracy in specific patient populations such as older patients or those with comorbidities.

Recently, we demonstrated that women differ from men in their sleep apnea phenotype and compensatory responses to upper airway obstruction during sleep.[10] Similarly, aging can also affect the sleep apnea phenotype, with older patients often having more central apneas due to frequent sleep disruptions and cardiovascular comorbidities.[11] In addition, there are sex differences in autonomic responses to respiratory events,[12] and aging has been reported to be associated with impairments in vascular tone, decreased overall heart rate variability, and impairment in cardiac control mechanisms,[13] all of which could influence the performance of WatchPAT in these populations.

The primary objectives of the current study were to develop and validate an approach for reviewing and editing automated WatchPAT scoring and to assess the accuracy of determining the severity of SDB in both an unselected clinical population and subpopulations of different ages and sexes. We therefore conducted WatchPAT recordings in parallel with in-lab PSG to optimize and validate an algorithm for editing automated WatchPAT scoring and compared WatchPAT automated and manual editing with standard PSG scoring. We hypothesized that visual editing of automated WatchPAT scoring would be reliable and accurate yet time-efficient.

METHODS

Participants

Patients over 18 years of age with suspected SDB were prospectively enrolled between January 2017 and September 2017. All patients, giving informed consent, wore WatchPAT 200 simultaneously with their in-lab PSG in a time-synchronization manner. Clinical and demographic data including comorbidity and medication were also collected. The study was approved by the Johns Hopkins Institutional Review Board on human research.

Sleep study

Polysomnography

Standard in-lab PSG was performed using a computerized polysomnographic device (Embla system, Flaga HF, Iceland), with the following channels: electrooculography, electrocardiography, electroencephalography (C2-A1, C3-A2, O2-A1, O1-A2, F4-A1, F3-A2), submental and tibial electromyography, airflow (thermistor, nasal pressure transducer), chest and abdominal motion, oxygen saturation (pulse oximetry), snoring microphone, and body position sensor. All the sleep studies were manually interpreted by sleep technicians according to the standard criteria, as outlined in *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*[14] and were reviewed by certified physicians. In particular, an apnea was scored if airflow was absent for 10 seconds and a hypopnea if airflow dropped by ≥ 30% of pre-event baseline in association with a desaturation of 3% or an arousal. Both the technician and the sleep physician were blinded to the WatchPAT signals to eliminate bias.

WatchPAT 200

Signals recorded by WatchPAT included the PAT signal (PAT probe), pulse rate (PAT signal), oxyhemoglobin saturation (pulse oximetry), wrist activity (actigraphy), and snoring (microphone). WatchPAT signals were imported into REMlogic for reviewing and editing. Each WatchPAT editing process was then entered into a database for developing an algorithm and testing its accuracy compared with PSG scoring, as outlined below in Study protocol. According to the automated algorithm, sleep/wake detection was based on assessment of movements and their occurrences (periodic or sporadic) while the sleep stage detections (rapid eye movement, non-REM [NREM] sleep) were based on the spectral and temporal components of the PAT signal.[15] A respiratory event was scored based on changes in PAT signal amplitude, pulse rate, and oxygen saturation as outlined below.

Manual algorithm for verifying and editing automated WatchPAT scoring

A process was developed to further refine scoring by editing WatchPAT automated sleep staging and respiratory events scoring as follows.

Manual algorithm for sleep staging

The amplitude and variability in pulse rate and PAT signals change with sleep stages as previously validated.[16,17] The visual inspection of these 2 signals allowed the user to check the plausibility of Wake and REM detection. The following procedure was adopted for reviewing the hypnogram in a 30- to 60-min time window.

1. WatchPAT's graphic display was reviewed and emblematic periods of Wake and REM were identified based on characteristics of Wake and REM sleep as detailed in **Table S1** in the supplemental material.
2. All Wake and REM periods were compared with aforementioned emblematic periods.
3. Sleep stages were corrected if PAT, pulse rate, and oximeter tracing did not show the typical features as outlined in **Table S1**. If a section of the recording did not definitely represent REM or Wake, it was revised to NREM sleep.

Manual algorithm for respiratory events scoring

Apneas and hypopneas are usually terminated with a sympathetic discharge. This discharge is characterized by a reciprocal pattern in PAT amplitude and pulse rate. This pattern consists of a decrease in PAT amplitude and concordant increase in pulse rate (as shown in **Figure S1** in the supplemental material). A development sample of patients' recordings was analyzed visually and yielded the following procedure for editing

respiratory events in a 10- to 15-min time window in a validation sample.

We developed and tested 2 algorithms (**Figure S2** and **Figure S3**):

Algorithm A: The criteria as following are applicable to both NREM and REM sleep.

1. SDB events were deleted if:
   i. The event did not have the typical pattern of PAT amplitude reduction coinciding with an increase in pulse rate.
   ii. The reciprocal pattern was associated with a positional change.
   iii. The reciprocal pattern was associated with a desaturation of < 3% or no change in snoring pattern (**Figure S2**).
2. SDB events were added if:
   A reciprocal pattern was detected visually along with a desaturation of ≥ 4% (**Figure S3**).

Algorithm B: The criteria mentioned above are only applicable to NREM sleep.

In REM sleep, SDB events were added or retained if any non-artifactual desaturation of ≥ 4% was present, regardless of whether a reciprocal pattern was also present.

Study protocol

The manual algorithm was optimized with 2 scorers (Z.Z. and H.S.) using a development set of 30 patients. To allow a wide range of SDB severities, the development set was comprised of 30 patients with evenly distributed WatchPAT oxygen desaturation index (ODI). Then the optimal algorithm was further tested with 2 scorers (Z.Z. and M.S.) using a training set of 62 patients. At last, the algorithm was validated with 3 scorer (Z.Z., H.S., and J.E) using another independent validation set of 170 patients meeting the same inclusion criteria.

Statistical analysis

Data were summarized as means ± standard deviation for continuous variables and as frequencies (percentage) for categorical variables. Spearman correlation coefficient and Bland-Altman plots were used to assess correlation and agreement for TST, REM sleep time, and AHI between PSG and WatchPAT, respectively. Receiver operator characteristic curve and area under the curve were assessed using an AHI cutoff of 5, 15, and 30 events/h on PSG and further compared using nonparametric and binormal methods to explore the diagnostic values of automated and manual scoring on WatchPAT. Sensitivity, specificity, likelihood ratios, and Youden index were calculated using 2 × 2 contingency tables. The Youden index, defined as "sensitivity + specificity − 1," was used to determine the optimal cut-off points in each analysis, where equal weight was given to the sensitivity and specificity of the test.

The Kendall τ-b statistic was used to evaluate concordance of severity of AHI between PSG and WatchPAT at cutoffs of 5, 10, 15, 20, 25, and 30 events/h, with concordance being excellent (0.80–1.00), very good (0.60–0.79), good (0.40–0.59), fair (0.20–0.39), and poor (0.00–0.19). Intraclass correlation coefficient was used to estimated interscorer reliability of manual scoring, with agreement being very strong (0.90–1.00), strong (0.70–0.89), moderate (0.5–0.69), low (0.26–0.49), and little (0.00–0.25).[18] Statistical analyses were performed using STATA Version 14.0 (College Station, TX). Differences were considered statistically significant at the 2-sided $P < .05$ level.

RESULTS

Participant characteristics and study flow

A total of 1,373 patients met eligibility criteria, of which 400 (29.1%) agreed to participate in this study and completed the simultaneous PSG and WatchPAT monitoring. Data from 138 participants were rejected because of WatchPAT total recording time (TRT) < 6 hours (n = 57), PSG total sleep time (TST) < 4 hours (n = 48), difference of TRT between WatchPAT and PSG > 1 hour (n = 63), proportion of total artifact time in TRT ≥ 0.3 or poor recording quality of WatchPAT or PSG recording (defined as a substantial portion being not able to be used to score sleep and respiratory events) (n = 36). Sixty-six cases met more than one rejection criteria. The final population for analysis (n = 262) had an age of 49.0 ± 14.8 years, body mass index of 36.1 ± 9.1 kg/m^2, and AHI of 24.3 ± 21.6 events/h and was divided into development set (n = 30), training set (n = 62), and validation set (n = 170) (**Figure 1**). **Table 1** shows the demographic and clinical characteristics of study participants in validation set. There were 77 men and 93 women, 26 participants with an age ≥ 65 years, and 144 with an age < 65 years. Our study sample did not include any patients with severe heart failure or decompensated lung or kidney disease as outlined by recent AASM contraindications for HSATs.[19]

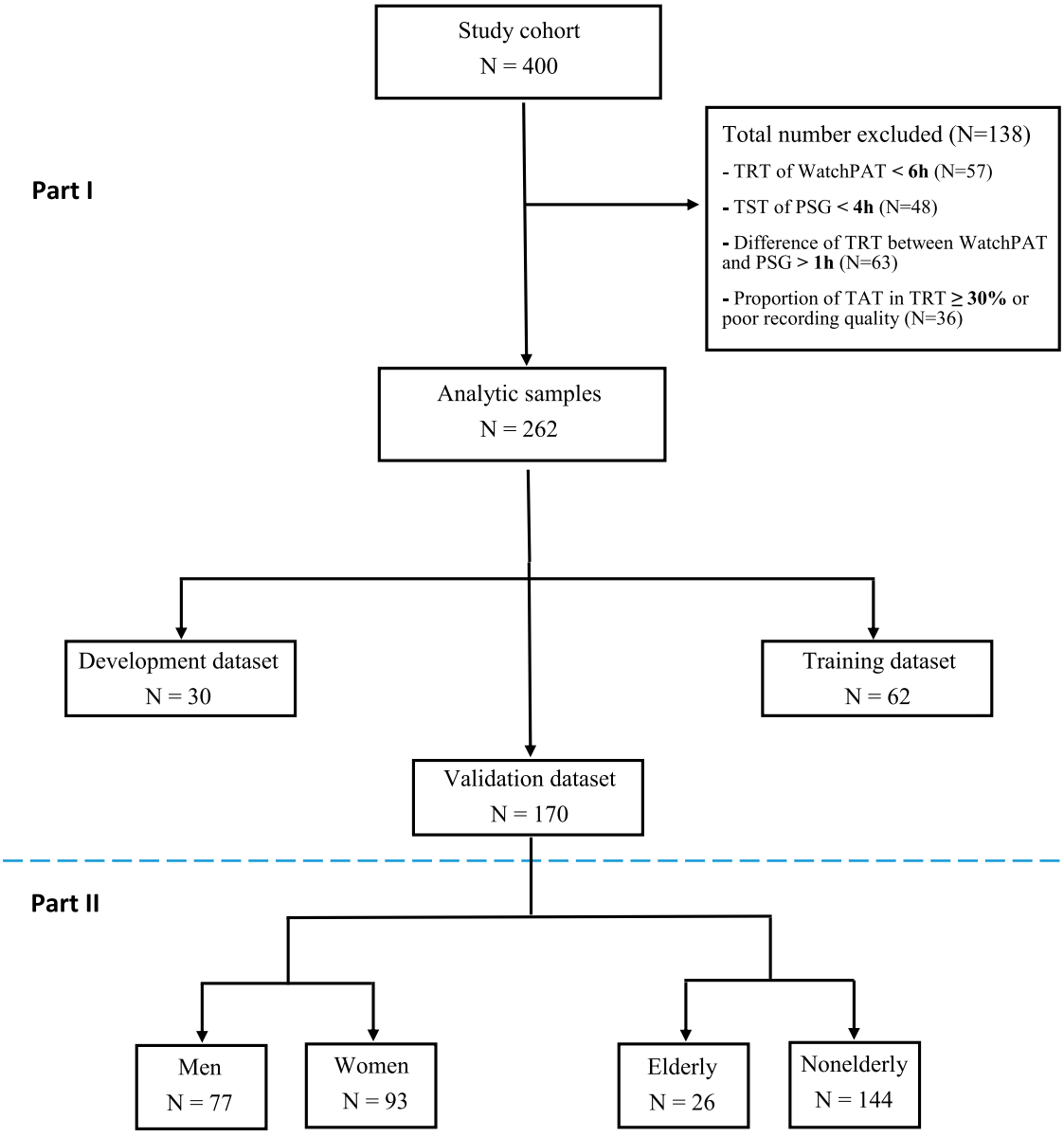
Development and validation of sleep time assessment

Table 2 shows that TST scored automatically was minimally different from PSG-derived TST. The manual editing of TST decreased the mean difference from 15.8 to 10.6 minutes at an average PSG-derived TST of 357.5 minutes with interscorer correlation of 0.94. The reference range of difference from PSG data and correlation with PSG data did not change substantially. In contrast, manual scoring of REM sleep time improved the correlation with PSG data and narrowed the reference range compared with automated scoring. The mean difference in REM sleep time decreased from 23.0 to 18.3 minutes at an average PSG-derived REM sleep time of 61.0 minutes. The interscorer correlation for manual REM sleep staging was 0.69. Thus, although estimation of TST was slightly affected by the editing process, estimation of REM sleep time was markedly improved, as confirmed in the Bland-Altman plots in **Figure 2**.

Development and validation of AHI editing algorithm

Table 3 presents the interscorer reliability of manually scored AHI from WatchPAT for the development, training, and validation set. It also shows the mean difference, reference range, and correlation of AHI between PSG and WatchPAT. First, in the development set, manual algorithm B demonstrated a higher correlation and better agreement with PSG than manual algorithm A. Second, using the larger training set, we found that algorithm B improved the correlation and agreement with PSG

Figure 1—Study protocol and flow diagram.



PSG = polysomnography, TAT = total artifact time, TRT = total recording time, TST = total sleep time.

compared with the automated algorithm and maintained the even higher interscorer correlation. Finally, in the validation set, the interscorer correlation was markedly higher than those in the development and training set. Although the automated scoring had a comparable AHI compared with PSG, the correlation was still rather low. In contrast, the manually edited AHI of all scorers slightly reduced the overall AHI but increased the correlation with the PSG-derived AHI and reduced the reference range, which is also shown in the Bland-Altman plots in **Figure 2**.

Manual scoring of WatchPAT significantly increased the area under the curve for classifying overall, moderate-severe, and severe SDB compared with automated scoring (0.858, 0.923, 0.909 and 0.766, 0.829, 0.831, respectively). (**Figure 3**) Sensitivity and specificity of diagnosis were better balanced and even improved with the original AHI cutoff values as well as the optimal AHI cutoff values (**Table S2**).

The concordance statistics for 2 categories of AHI severity (AHI cutoff value = 5, 10, 15, 20, 25, and 30 events/h, respectively) are shown in **Figure 4**. Compared with automated scoring of WatchPAT, manual scoring had higher concordances at all AHI cutoff values, particularly at cutoff of 5, 20, 25 and 30 events/h.

The manual editing of respiratory events in REM sleep by all scorers substantially reduced the mean difference of REM AHI from 5.1 to −2.8 events/h, with an interscorer correlation of 0.86. At cutoff values for REM AHI of 5 and 15 events/h, no significant differences emerged in concordance between automated and manual scoring algorithm. In contrast, the concordance for detecting severe REM-related sleep apnea

Table 1—Demographic and clinical characteristics for study participants in validation set.

Demographic/Clinical Characteristics	Mean ± SD or n (%)
Age, y	48.4 ± 14.9
≥ 65 y	26 (15.3)
Men	77 (45.3)
BMI (kg/m^2)	35.5 ± 9.3
AHI (events/h)	20.2 ± 18.9
≥5	135 (79.4)
≥15	82 (48.2)
≥30	43 (25.3)
Comorbidity	
Hypertension	59 (34.7)
Diabetes	38 (22.4)
Arrhythmia	
Atrial fibrillation	4 (2.4)
Atrial flutter	2 (1.2)
Tachy-brady syndrome	1 (0.6)
Premature atrial contractions	1 (0.6)
Premature ventricular contractions	3 (1.8)
Pacemaker	1 (0.6)
Insomnia	15 (8.8)
Restless legs syndrome/Periodic limb movements	8 (4.7)
Coronary artery disease	13 (7.6)
Heart failure	6 (3.5)
Chronic obstructive pulmonary disease	8 (4.7)
Cerebrovascular disease	6 (3.5)
Depression	2 (1.2)
Peripheral vascular disease	2 (1.2)
Circadian rhythm sleep disorder	6 (3.5)
Hypothyroidism	8 (4.7)
Medication	
Beta-blocker	15 (8.8)
α1-Adrenergic antagonist	3 (1.8)
Opiates	11 (6.5)
Antidepressant	22 (12.9)
Angiotensin-converting enzyme inhibitor	13 (7.6)
Angiotensin receptor blocker	10 (5.9)
Calcium channel blocker	7 (4.1)
Diltiazem	3 (1.8)
Diuretic	15 (8.8)
Benzodiazepine	8 (4.7)
Antidiabetic	12 (7.1)
Levothyroxine	8 (4.7)

Data are presented as means ± standard deviation. AHI = apnea-hypopnea index, BMI = body mass index.

at a REM AHI cutoff of 30 events/h increased with manual editing from 0.44 to 0.67.

Effect of age and sex on WatchPAT performance

The effects of age and sex on automatically and manually derived AHI are shown in **Figure 5**. For the automated algorithm, the concordance statistics in the older participants were fair to good. The manual algorithm markedly improved the concordances at most AHI cutoff values to a good to very good concordance, although there was no difference at the cutoff value of 30 events/h (**Figure 5**, top right). In the younger participants, AHI concordance for the automated algorithm was already good for most cutoff values and the manual editing increased the concordance to very good for all but the AHI cutoff of 5 events/h (**Figure 5**, top left). A similar difference was also observed between sexes. Although the automated algorithm performed similarly between men and women, the impact of manual editing was greater in women than that in men (**Figure 5**, bottom).

DISCUSSION

In comparing the results of automated and manual WatchPAT scoring with PSG in an unselected clinical population as well as age and sex substrata, we generated several novel findings. First, our editing algorithm for sleep staging and respiratory events scoring generated moderate to very strong interscorer reliabilities. Second, although the estimation of TST was comparable between automated and manual algorithm, the estimation of REM sleep time was markedly improved by manual editing. Third, manual scoring improved the correlation and agreement with PSG-derived AHI and the concordance statistic for categorical agreement for all AHI thresholds. Fourth, automated algorithms performed better in younger than in older participants, while it performed similarly between men and women with respect to concordance statistics. Manual algorithms markedly improved concordances more in older participants and women than in their counterparts. Our findings lead us to conclude that manual editing of WatchPAT scoring is reliable and improves the agreement with PSG-derived sleep and apnea and hypopnea indices.

Algorithm development

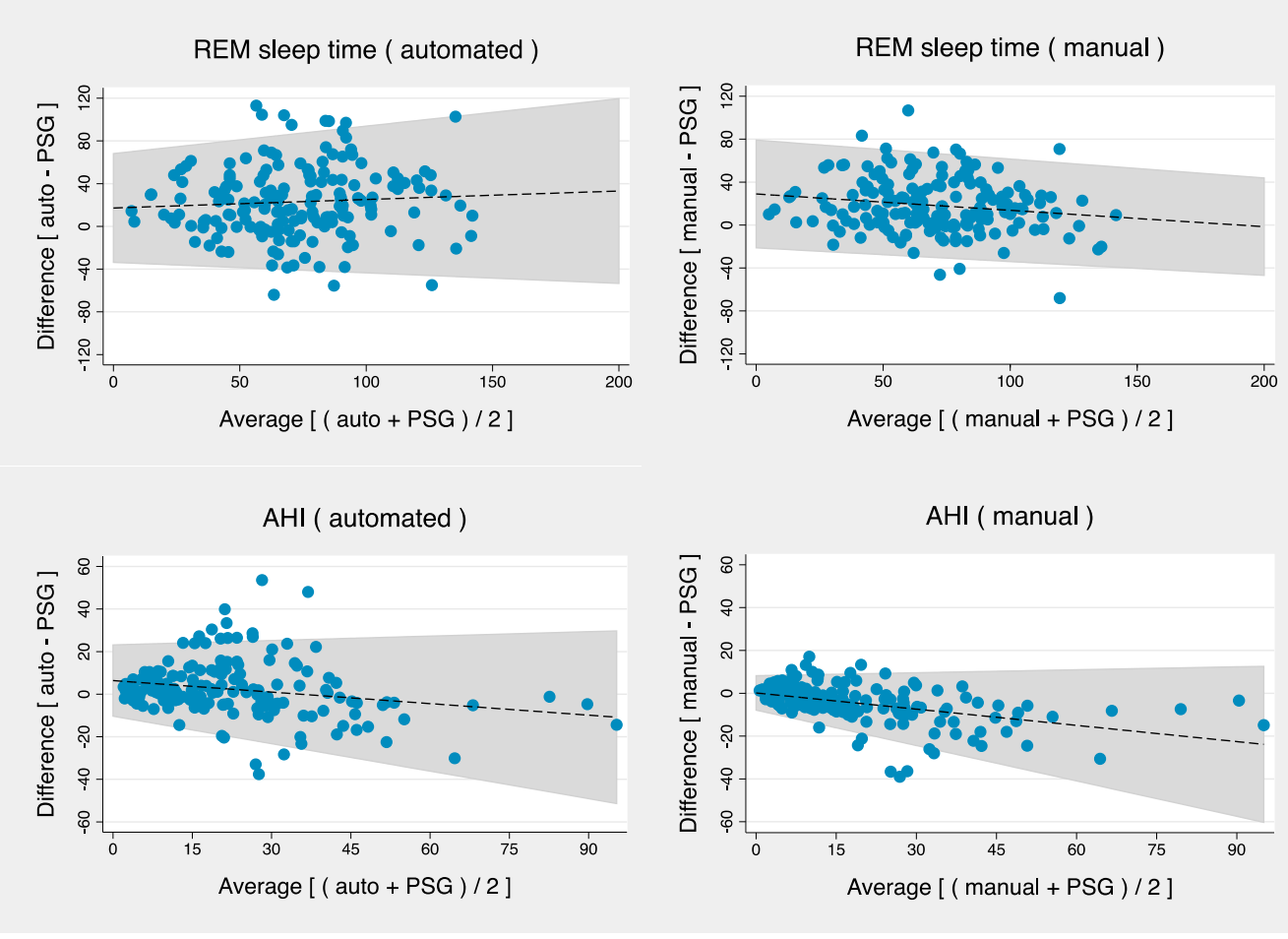
Over the years, WatchPAT has become a well-validated, widely used HSAT device. Despite extensive validation of its automated algorithms for sleep staging and sleep apnea scoring, the AASM requires physicians to review and evaluate the accuracy of sleep study reports.[20] We adopted a pragmatic approach to facilitate this process. First, we leveraged time efficiencies from automated scoring to streamline the editing process. Visual oversight yielded modest improvements in estimations of TST and REM sleep time compared with PSG. Second, SDB events were critically reviewed while imposing specific criteria for keeping, deleting, or adding these events. Our approach emphasized the recognition of sympathetic responses ("reciprocal pattern") to respiratory disturbances including oxyhemoglobin desaturation and snoring while rejecting responses that were

Table 2—Agreement and correlation of TST and REM sleep time with PSG and interscorer reliability in validation set.

Algorithm	Scorer	Total Sleep Time, min			REM Sleep Time, min		
		ICC	Agreement	Spearman's correlation	ICC	Agreement	Spearman's correlation
Manual (B)	R1	0.94	7.3 (−64.3–79.0)	.73	0.69	15.1 (−44.2–74.4)	.53
	R2		9.4 (−63.4–82.3)	.72		13.7 (−38.1–65.6)	.68
	R4		15.2 (−7.6–87.9)	.72		26.0 (−37.4–89.4)	.52
	Manual[a]		10.6 (−58.9–80.2)	.73		18.3 (−32.6–69.1)	.64
Automated			15.8 (−57.7–89.2)	.70		23.0 (−43.9–89.8)	.48

Data are presented as mean difference (reference range). [a]Average of R1, R2, and R4. ICC= intraclass correlation coefficient, PSG = polysomnography, REM = rapid eye movement. TST = total sleep time.

Figure 2—Bland–Altman plots for automated and manual scoring of REM sleep time and AHI in validation set.



For REM sleep time, manual scoring (top right) reduced the mean difference (dashed line) from PSG and narrowed the reference range (shaded area) compared with automated scoring (top left). For AHI, manual scoring (bottom right) also reduced the reference range (shaded area) compared with automated scoring (bottom left). AHI = apnea-hypopnea index, auto = automated, PSG = polysomnography, REM = rapid eye movement.

related to movement artifacts. Our algorithm was specifically designed to detect apnea and hypopnea events that were characterized by physiologic markers for desaturation and arousal, rather than respiratory effort-related arousals with marginal alterations in these markers. This approach led to significant improvements in classifying patients along the spectrum from mild, moderate, and severe sleep apnea compared with automated WatchPAT results. In circumventing de novo scoring of the native recording, we were able to streamline the visual editing of automated scores and improve agreement with gold standard PSG.

Estimation of sleep staging

We also assessed the accuracy of detecting TST and REM sleep time in the current study. Whereas the automated detection of TST is based on actigraphy, the REM detection is based on a

Table 3—Agreement and correlation of AHI with PSG and interscorer reliability based on automated or manual algorithm.

Data Set	Algorithm	Scorer	ICC	Agreement	Spearman's Correlation
Development (n = 30)	Automated			5.4 (−20.9 to 31.7)	.79
	Manual (A)	R1	0.84	−6.8 (−29.1 to 15.5)	.86
		R2		−4.6 (−30.9 to 21.7)	.83
	Manual (B)	R1	0.87	−6.0 (−27.7 to 15.7)	.86
		R2		−3.7 (−27.0 to 19.7)	.84
Training (n = 62)	Automated			5.4 (−26.0 to 36.7)	.75
	Manual (B)	R1	0.93	−3.6 (−24.5 to 17.4)	.88
		R3		2.9 (−25.3 to 31.0)	.78
Validation (n = 170)	Automated			2.5 (−24.0 to 28.9)	.65
	Manual (B)	R1	0.96	−4.6 (−22.8 to 13.6)	.81
		R2		−4.1 (−24.6 to 16.5)	.77
		R4		−4.7 (−22.0 to 12.6)	.84
		Manualb		−4.5 (−22.5 to 13.6)	.81

Data are presented as mean difference (reference range). [b]Average of R1, R2, and R4. AHI = apnea-hypopnea index, ICC = intraclass correlation coefficient, PSG = polysomnography.

Figure 3—Receiver operator characteristic (ROC) curves using an apnea-hypopnea index (AHI) cutoff of 5, 15, and 30 events/h on the polysomnography (PSG) (top, middle, and bottom, respectively).

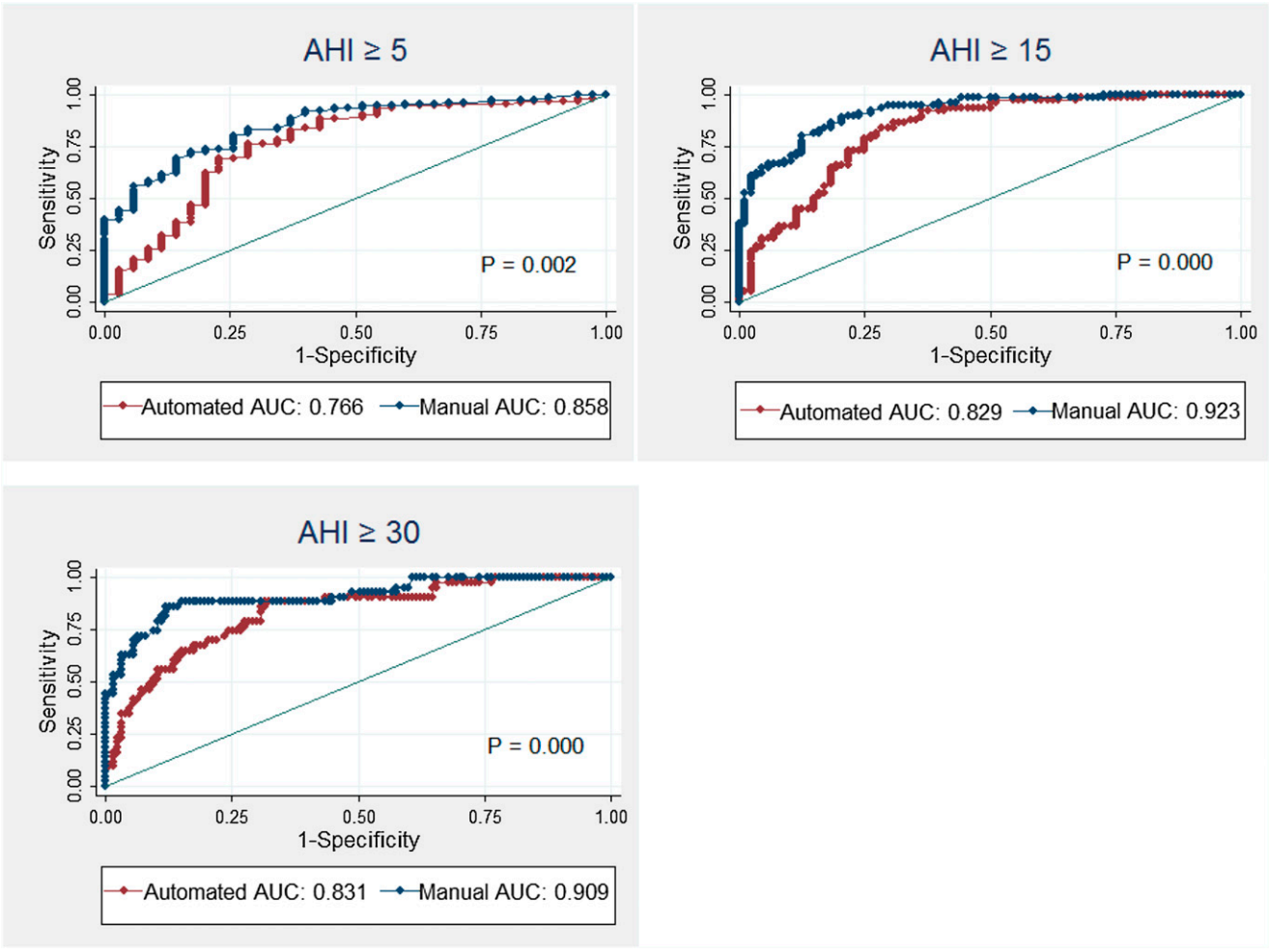
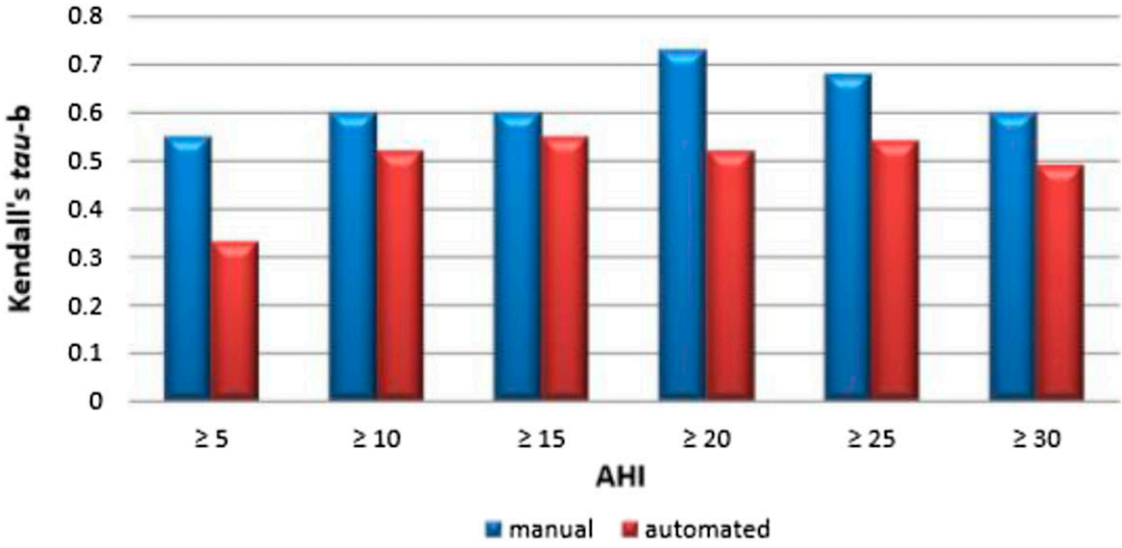


Figure 4—Concordance between PSG and WatchPAT in the validation set at specific AHI cutoff values.



Compared with automated scoring of WatchPAT, manual scoring had higher concordances at all AHI cutoff values, particularly at cutoff values of 5, 20, 25, and 30 events/h. AHI = apnea-hypopnea index, PSG = polysomnography.

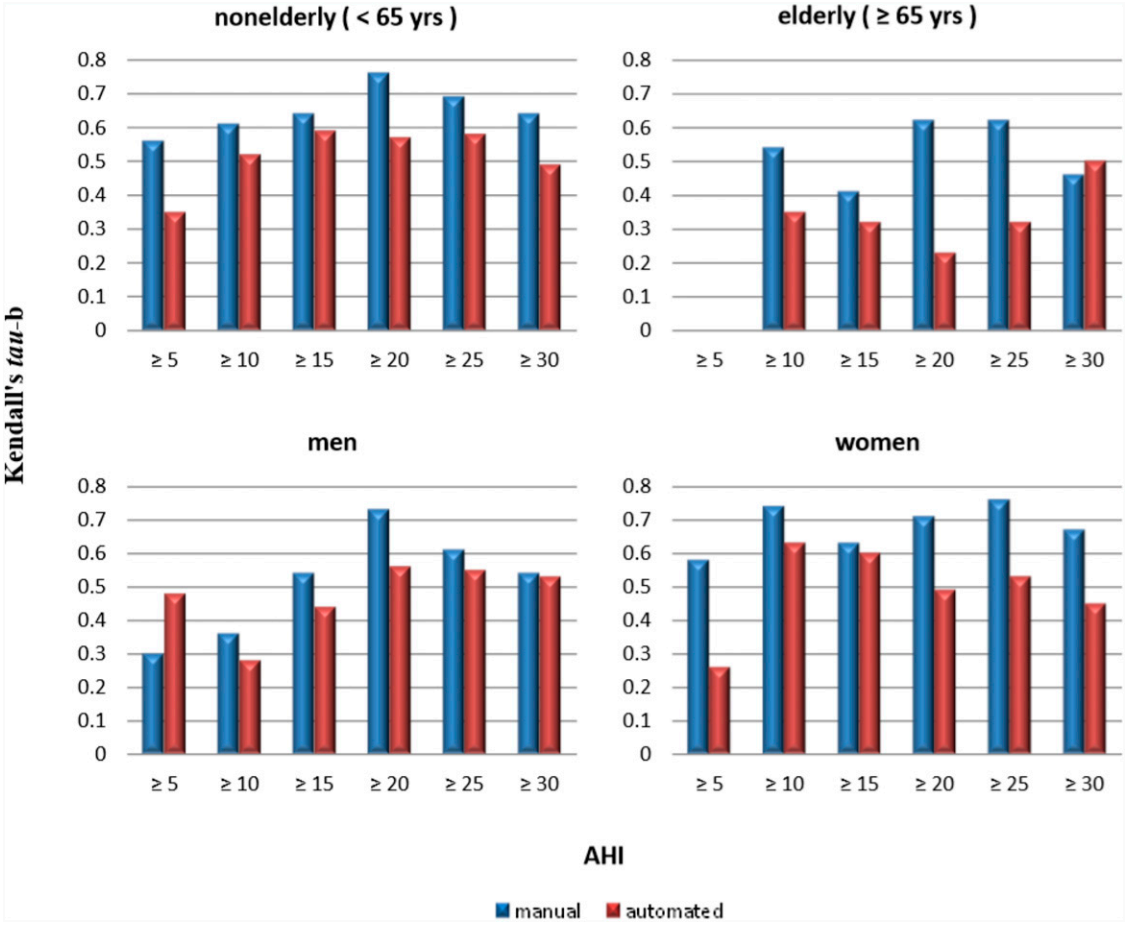
number of features derived from spectral and temporal components of the heart rate and PAT signal.[15] For the reviewing and editing of TST and REM sleep time, we inspected all WatchPAT signals, including snoring, oxygen saturation, actigraphy, PAT, and pulse rate. Although the automated estimation of TST already differed minimally (+15.8 minutes with a mean sleep time of 357.5 minutes: + 4.4%), the editing process lowered the TST difference by another 5.2 minutes, decreasing any overestimation of TST to less than 3%. Estimation of REM sleep time was also improved by about 5 minutes with the editing process. Although the interscorer correlation was only moderate, the result was essentially within the variability range reported in some studies comparing registered PSG scorers[21] and similar to the agreement reported in comparisons between automated PSG scoring and manual scoring.[22] Given the average REM sleep time of approximately 1 hour, this improvement had a more significant impact on accuracy. Thus, estimations of TST and REM sleep time were well within the acceptable error rate for determining these parameters on PSG.

Estimation of AHI

Our method for recognizing respiratory events is based on detecting a visible arousal response as defined by the presence of both a substantial reduction in PAT amplitude and a rise in pulse rate (which we call a reciprocal pattern). This response often terminates apneic and hypopneic events during NREM sleep and remains an initial pivotal feature required for confirming or rejecting events in NREM sleep. We discovered that this reciprocal pattern was less pronounced, however, during REM sleep due to an overall reduction in PAT amplitude in this sleep stage ("floor effect"). Nonetheless, the sensitivity in detecting respiratory events in REM sleep was maintained by keeping automatically scored events and adding events whenever accompanied by a 4% desaturation, even in the absence of the reciprocal pattern. This modification (algorithm B) achieved a higher correlation and better agreement with AHI on PSG than both algorithm A and the automated algorithm. In the validation set, we further demonstrated that algorithm B remained reliable and valid, yet simple to execute. More importantly, compared with automated scoring of WatchPAT, manual scoring with algorithm B produced better concordances across the entire range of AHI severity cutoffs. Thus, visual editing of the automated WatchPAT scoring improved the AHI concordance with PSG substantially. Moreover, the significance was also highlighted by the determination of severe REM related sleep apnea (REM AHI ≥ 30 events/h), which had a very good concordance of 0.67 and a strong interscorer correlation of 0.86. Thus, even the detection of severe REM-related sleep apnea was within standard margins of error for AHI in REM sleep.

It is noteworthy that, although PSG is a standard for sleep testing, it is not a perfect test in several ways. PSG scoring itself is subject to major variability, night-to-night in the same center, across sleep centers, and between scorers. In our study, only a single PSG night and WatchPAT night study were compared. PSG scoring included hypopneas of ≥ 3% desaturation and ≥ 30% NC pressure drop. This definition may score "physiologic" hypopneas compared with possibly pathological hypopneas associated with autonomic activation per PAT. It would be worthwhile to compare the variability of AHI between 2 nights of PSG and 2 if not even 3 nights of WatchPAT studies in the home, particularly in patients with mild sleep apnea in whom the night to night variability may then affect the disease category. In addition, such a study would not only show the value repeated sleep studies for making the diagnosis, but also demonstrate the value of single-night PSG.

Figure 5—Concordance between PSG and WatchPAT at specific AHI cutoff values in subpopulations of different age and sex.



The automated algorithm performed better in younger patients (top left) than in older patients (top right), while it performed similarly in men (bottom left) and women (bottom right) with respect to concordance statistics. The manual algorithm markedly improved concordance in the older patients and in the women compared with younger patient and men subgroups. AHI = apnea-hypopnea index, PSG = polysomnography.

Effect of sex and age on performance (unselected clinical population)

We supposed that the algorithms to detect sleep apnea would work better in men and in younger patients due to the preponderance of men and younger participants in previous validation studies. Our clinical population consisted of a high percentage of women (55%), thus allowing us to assess the effect of sex on scoring concordance. Strikingly, we found that the performance of the automated algorithm was similar between women and men, yet the manual editing process improved the concordance more in women than in men. In contrast, younger patients had better concordances for the automated and manual algorithm than older patients. Compared with automated algorithm, manual editing markedly improved the concordances in the older patients at most AHI cutoff values to a good to very good concordance, which becomes particularly relevant for testing an aging population. The reasons for sex and age differences in concordance may be related to differences in vascular compliance and homeostasis due to hormonal, neural, or anatomic differences as well as increased comorbidities and medication use in older patients.[10–13]

We, however, recognize that there was a paucity of older patients, with only 25 (15%), and thus it is more difficult to generalize about accuracy of scoring methods, especially when it comes to the subanalysis by AHI severity. We, however, believe that our current subanalysis will build the foundation for future validation studies of HSATs in an older patient population. Nevertheless, women and older patients often exhibit different SDB patterns that may have further influenced the performance of determining respiratory events either by automated or manual editing.

Strengths and weaknesses

One of the strengths of this study is that it relied on a large, unselected, sleep clinic-based cohort with PSG recordings using up-to-date recording techniques. Recordings were analyzed manually by board-certified technicians and physicians, which will likely increase its applicability and generalizability in clinical real-world settings. Moreover, both technicians and physicians were blinded to WatchPAT signals, which helped reduce bias. Another advantage was that we validated an independent data set, and demonstrated high efficiency,

reliability, and validity of the manual algorithm. Although the mean differences shown between TST, REM scoring of WatchPAT auto versus manual versus PSG are small, we identified several studies that demonstrated either a marked overestimation or underestimation of the automated AHI versus the PSG or the manual edited AHI. We did not find a systematic reason for this discrepancy. With visual editing, we markedly narrowed the reference range to −22.5 to 13.6 as shown in **Table 3** and **Figure 2** and improved the concordance with PSG for categorical agreement of SDB severity. In conclusion, the user of the WatchPAT must be aware that outliers exist and that manual editing can identify and correct these reports adequately.

Nevertheless, some limitations of the study also deserve mention. First, PSGs were not double-scored, thereby preventing us from calculating interscorer correlation and an epoch by epoch comparison between PSG and WatchPAT results. Yet our technicians and physicians demonstrated consistent scoring performance on standard monthly interscorer reliability exercises in our laboratory, and the scoring approach modeled that practiced in the clinical setting. Second, we recognize that many patients were excluded from analysis in our study to compare recordings from WatchPAT with PSG. We did not a priori exclude patients with a primary complaint of difficulties initiating or maintaining sleep. There were 25 patients that slept < 3 hours and 48 patients whose overall sleep time was between 3 and 4 hours. The WatchPAT automated algorithm requires a sleep time of > 3 hours and 30 minutes of REM sleep for calculating NREM/REM distribution and the AHI during NREM and REM accordingly. Thus, participants were excluded post hoc as we could not compare results of the automated analysis with the edited and PSG indices. Third, our study is a real-world study and we want to develop a new algorithm applicable to clinical work no matter if the patient is OSA-dominant or central sleep apnea-dominant. We therefore refer to "SDB" or "sleep apnea" from the beginning to the end instead of "OSA." Future studies are required to validate if the WatchPAT can accurately differentiate between central and obstructive sleep-disordered events. Fourth, despite marked improvements in overall accuracy of WatchPAT sleep and AHI indices, ~10% of the studies showed either a marked overestimation or underestimation of AHI compared with PSG scoring. Nevertheless, all of these studies could be identified by experts trained to review WatchPAT signals. Fifth, our manual method was based on recognizing the sympathetic response to respiratory events, as a marker for cortical and subcortical "arousals," although AASM criteria only use cortical arousals to define hypopneas.

Implications

Our findings have several clinical and research implications. First, WatchPAT provides accurate estimations of TST and REM sleep time. SDB events are detected based on either a significant oxygen desaturation or a visible arousal response, both of which help characterize acute pathophysiologic responses to SDB episodes while improving the diagnostic accuracy in determining AHI. Second, the estimations of TST and REM sleep time also help characterize disturbances in sleep architecture that may be independent of sleep apneic activity. Third, REM sleep recognition makes it possible to identify patients with predominant REM-related sleep apnea, which may be important in risk-stratifying patients prior to surgical procedures or in determining the effect of REM-related sleep apnea on neurocognitive and cardiovascular function. Finally, the time for reviewing and editing the hypnogram and AHI of the WatchPAT averaged ~10–15 minutes. Thus, the time efficiency and improvements in accuracy promise to reduce operational costs in larger clinical or pharmacological research studies. Thus, our data demonstrate that manual editing of WatchPAT automated scoring is reliable and improves agreement with PSG-derived sleep and apnea and hypopnea indices across age and sex strata.

ABBREVIATIONS

AASM, American Academy of Sleep Medicine
AHI, apnea-hypopnea index
HSAT, home sleep apnea test
NREM, non-rapid eye movement
ODI, oxygen desaturation index
PAT, peripheral arterial tone
PSG, polysomnography
REM, rapid eye movement
SDB, sleep-disordered breathing
TRT, total recording time
TST, total sleep time

REFERENCES

1. Peppard PE, Young T, Palta M, Skatrud J. Prospective study of the association between sleep-disordered breathing and hypertension. *N Engl J Med*. 2000;342(19):1378–1384.
2. Yaggi HK, Concato J, Kernan WN, Lichtman JH, Brass LM, Mohsenin V. Obstructive sleep apnea as a risk factor for stroke and death. *N Engl J Med*. 2005;353(19):2034–2041.
3. Whitaker KM, Lutsey PL, Ogilvie RP, et al. Associations between polysomnography and actigraphy-based sleep indices and glycemic control among those with and without type 2 diabetes: The Multi-Ethnic Study of Atherosclerosis. *Sleep*. 2018;41(11):zsy172.
4. Campos-Rodriguez F, Martinez-Garcia MA, Martinez M, et al. Spanish Sleep Network. Association between obstructive sleep apnea and cancer incidence in a large multicenter Spanish cohort. *Am J Respir Crit Care Med*. 2013;187(1):99–105.
5. Zou D, Grote L, Peker Y, Lindblad U, Hedner J. Validation a portable monitoring device for sleep apnea diagnosis in a population based cohort using synchronized home polysomnography. *Sleep*. 2006;29(3):367–374.
6. O'Brien LM, Bullough AS, Shelgikar AV, Chames MC, Armitage R, Chervin RD. Validation of Watch-Pat-200 against polysomnography during pregnancy. *J Clin Sleep Med*. 2012;8(3):287–294.
7. Park CY, Hong JH, Lee JH, et al. Clinical usefulness of watch-PAT for assessing the surgical results of obstructive sleep apnea syndrome. *J Clin Sleep Med*. 2014;10(1):43–47.
8. Gan YJ, Lim L, Chong YK. Validation study of WatchPat 200 for diagnosis of OSA in an Asian cohort. *Eur Arch Otorhinolaryngol*. 2017;274(3):1741–1745.
9. Yalamanchali S, Farajian V, Hamilton C, Pott TR, Samuelson CG, Friedman M. Diagnosis of obstructive sleep apnea by peripheral arterial tonometry: meta-analysis. *JAMA Otolaryngol Head Neck Surg*. 2013;139(12):1343–1350.

10. Chin CH, Kirkness JP, Patil SP, et al. Compensatory responses to upper airway obstruction in obese apneic men and women. *J Appl Physiol*. 2012;112(3):403–410.
11. Bixler EO, Vgontzas AN, Ten Have T, Tyson K, Kales A. Effects of age on sleep apnea in men: I. Prevalence and severity. *Am J Respir Crit Care Med*. 1998;157(1):144–148.
12. Wadhwa H, Gradinaru C, Gates GJ, Badr MS, Mateika JH. Impact of intermittent hypoxia on long-term facilitation of minute ventilation and heart rate variability in men and women: do sex differences exist? *J Appl Physiol*. 2008;104(6):1625–1633.
13. Itzhaki S, Lavie L, Pillar G, Tal G, Lavie P. Endothelial dysfunction in obstructive sleep apnea measured by peripheral arterial tone response in the finger to reactive hyperemia. *Sleep*. 2005;28(5):594–600.
14. Berry RB, Brooks R, Gamaldo CE, et al. for the American Academy of Sleep Medicine. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*.Version 2.3 Darien, IL: American Academy of Sleep Medicine; 2016.
15. Hedner J, White DP, Malhotra A, et al. Sleep staging based on autonomic signals: a multi-center validation study. *J Clin Sleep Med*. 2011;7(3): 301–306.
16. Lavie P, Schnall RP, Sheffy J, Shlitner A. Peripheral vasoconstriction during REM sleep detected by a new plethysmographic method. *Nat Med*. 2000;6(6):606.
17. Somers VK, Dyken ME, Mark AL, Abboud FM. Sympathetic-nerve activity during sleep in normal subjects. *N Engl J Med*. 1993;328(5): 303–307.
18. Plichta SB, Kelvin EA. *Munro's Statistical Methods for Health Care Research*. 5th ed. Philadelphia, PA: Lippincott Williams Wilkins; 2005.
19. Kapur VK, Auckley DH, Chowdhuri S, et al. Clinical practice guideline for diagnostic testing for adult obstructive sleep apnea: an American Academy of Sleep Medicine clinical practice guideline. *J Clin Sleep Med*. 2017;13(3):479–504.
20. Collop NA, Anderson WM, Boehlecke B, et al. Clinical guidelines for the use of unattended portable monitors in the diagnosis of obstructive sleep apnea in adult patients. *J Clin Sleep Med*. 2007;3(7):737–747.
21. Norman RG, Pal I, Stewart C, Walsleben JA, Rapoport DM. Interobserver agreement among sleep scorers from different centers in a large dataset. *Sleep*. 2000;23(7):901–908.
22. Svetnik V, Ma J, Soper KA, et al. Evaluation of automated and semi-automated scoring of polysomnographic recordings from a clinical trial using zolpidem in the treatment of insomnia. *Sleep*. 2007;30(11):1562–1574.

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication June 9, 2019
Submitted in final revised form November 27, 2019
Accepted for publication November 27, 2019
Address correspondence to: Hartmut Schneider, MD, Johns Hopkins Sleep Disorders Center, Division of Pulmonary and Critical Care Medicine, 5501 Hopkins Bayview Circle, Baltimore, MD 21224; Email: hschnei3@gmail.com

DISCLOSURE STATEMENT

All authors have seen and approved the manuscript. Work for this study was performed at Johns Hopkins Sleep Disorders Center. The study is a result of a project funded by Itamar Medical Ltd., which had no access to the results of the study. Itamar Medical LTD, Caesaria, Israel, provided financial support for the study for the years 2016 to 2018. The terms of these arrangements were being managed by the Johns Hopkins University in accordance with its conflict of interest policies. In addition, from the years 2018 to 2019, authors H. Schneider and A. Schwartz received consulting fees from Itamar Inc. Atlanta, GA, for continued educational lectures of how to review and edit WatchPAT recordings. All other authors report no conflicts of interest.