PHILIPS

RESPIRONICS

Somnolyzer

Validation of the new Somnolyzer automated scoring system against manual scoring of respiratory events, arousals, periodic limb movements and sleep staging

Contributors: Peter Anderer, Jessie Bakker, Andreas Cerny, Bill Hardy, Jeff Jasko, Marco Ross, Edmund Shaw, Ray Vasko, David White.[1]

[1] Contributors were employees of Philips at the time this work was completed.

Introduction

Moderate-to-severe obstructive sleep apnea (OSA) is estimated to affect as much as 14.5% of the adult population of the United States, 8.8% in China, 14% in Japan and between 12% to over 30% in many countries in Western Europe[1] by current estimates, up to 80% of those with OSA are undiagnosed,[2] suggesting a need to expand accessibility to diagnostic testing. Interpretation of a polysomnography (PSG), the gold-standard test for diagnosing OSA, requires sleep staging as well as identification of breathing events, arousals and limb movements, in order to generate metrics such as the apnea-hypopnea index (AHI). Ideally, the results of PSG data reviews should be consistent among technologists and facilities; however, scoring variability remains an issue[3,4] and is further challenged by the need for more efficient reviews of sleep studies in many clinical settings.

Automated PSG scoring systems have been available for some time[5] although, until recently, a technologist review of the auto-scoring data was required before final review and interpretation by a physician. In October 2020, the US Food and Drug Administration (FDA) cleared for marketing the Philips Somnolyzer 4.1 auto-scoring solution as providing “physician-ready” output, meaning that adjustment of the scoring by a technologist is not required. The results of four clinical studies were provided to the FDA to make their assessment, three of which compared Somnolyzer to multiple human scorers. The purpose of the final study, presented here, was to demonstrate the robustness of algorithm performance when applied to data drawn from the National Sleep Research Resource (NSRR),[6,7] a publicly available dataset of sleep studies, each scored by a single technologist. The over-arching objective was to assess the agreement between Somnolyzer and manual sleep staging and event scoring, as described in Box 1 below.

Box 1

Objectives	Hypotheses
To determine the agreement between Somnolyzer sleep staging and manual sleep staging.	The lower-bound of the 95% CI of Cohen’s kappa based on epoch-by-epoch comparisons of all sleep stages (W/N1/N2/N3/R) between Somnolyzer and manual scoring will exceed 0.60.
To determine the agreement between the Somnolyzer-AHI and manual-AHI.	The lower-bound of the 95% CI of the ICC between the Somnolyzer-AHI and manual-AHI will exceed 0.75.
To determine the agreement between the Somnolyzer-ArI and manual-ArI.	The lower-bound of the 95% CI of the ICC between the Somnolyzer-ArI and manual-ArI will exceed 0.5.
To determine the agreement between the Somnolyzer-PLMSI and manual-PLMSI.	The lower-bound of the 95% CI of the ICC between the Somnolyzer-PLMSI and manual-PLMSI will exceed 0.75.

AHI = apnea hypopnea index
ArI = arousal index
PLMSI = periodic limb movements of sleep index
CI = confidence interval
ICC = intraclass correlation coefficient
W = wake
R = rapid eye movement

Methods

This study was approved by the Western Institutional Review Board (20192296). All PSG data were de-identified and therefore the requirement for informed consent was waived.

PSG identification and scoring

The PSGs used in this study were originally collected in the Apnea, Bariatric Surgery, and CPAP trial (ABC; n=24),[8] the HomePAP trial (n=180)[9] and the Multi-Ethnic Study of Atherosclerosis study (MESA; n=224).[10] After applying parameters to ensure a wide range of disease severity and to ensure a recording of ≥4 hours per study, PSGs were selected from each dataset at random.

The original sleep staging and event identification were left unchanged from the manual scoring already performed for each study following American Academy of Sleep Medicine (AASM) recommendations. For the MESA and ABC studies, hypopneas were identified when associated with a ≥3% SpO_2 desaturation and/or an arousal. In the HomePAP study, hypopneas were identified when associated with a ≥4% SpO_2 desaturation. Limb movements data based on leg-EMG were available only in the HomePAP study. Each PSG was analyzed in Sleepware G3 software containing Somnolyzer 4.1 (Philips, Monroeville, PA USA) following the original scoring criteria after applying the same lights off/on times.

Statistical power

An *a priori* power calculation was undertaken for all four endpoints described in our hypotheses above. The study was powered based on the weakest effect size (ICC between the Somnolyzer-ArI and manual-ArI). For that endpoint, a sample of n=425 was required.

Statistical analyses

All analyses were performed using Matlab R2019b, validated against IBM SPSS (Version 19.0.0.2). To assess sleep staging performance, we undertook an epoch-by-epoch comparison of manual staging and Somnolyzer staging and calculated a kappa statistic across all sleep stages (W/N1/N2/N3/R), as well as each individual stage, along with 95% CIs. We also calculated accuracy for each sleep stage discrimination; that is, the percentage of all epochs that were correctly identified by Somnolyzer. To assess the performance of respiratory event, arousal and limb movement identification, we computed the ICC for absolute agreement between Somnolyzer and manual scoring of the AHI, ArI and PLMSI, along with a 95% CI.

In addition to the performance targets adopted in our hypotheses, we compared the lower limit of the kappa 95% CI values to the thresholds defined by Landis and Koch (1977)[7] as follows: 0.0-0.2 slight agreement; 0.21-0.40 fair agreement; 0.41-0.60 moderate agreement; 0.61-0.80 substantial agreement; 0.81-1.0 almost perfect or perfect agreement. The lower limits of the ICC 95% CIs were compared against the thresholds defined by Koo and Li (2016)[11] as follows: <0.5 poor reliability; 0.5 to <0.75 moderate reliability, 0.75 to <0.90 good reliability; ≥0.90 excellent reliability.

Results

A total of 428 PSGs were randomly selected from within disease severity categories (AHI<5, 5 to <15, 15 to <30 and ≥30 events/hour). A small number were removed due to invalid signals (complete or partial) or missing manual scoring. The final sample sizes for each hypothesis were therefore: n=426 (sleep staging and AHI), n=425 (ArI) and n=174 (PLMSI; based on HomePAP data only). Descriptive information is provided in Table 1. As anticipated, with the sampling strategy and the nature of each study, each sample contained participants with a wide range of disease severity (AHIs in MESA 0 to 88 events/hour; HomePAP 0 to 113 events/hour; ABC 15 to 115 events/hour).

Table 1: Descriptive demographic and clinical information

Variable	MESA (n=224)	HomePAP (n=178)	ABC (n=24)
Age (years)	69.8±8.8	46.1±12.1	50.5±9.1
Gender (number; %)			
Female	110; 49.1%	89; 50.0%	14; 58.3%
Male	114; 50.9%	89; 50.0%	10; 41.7%
Ethnicity (number; %)	–		
Hispanic		14; 7.9%	2; 8.3%
Non-Hispanic		163; 91.6%	22; 91.7%
Unknown		1; 0.6%	–
Race (number; %)	–		
White		124; 69.7%	16; 66.7%
African-American		44; 2.7%	4; 16.7%
Other		10; 5.6%	16.7%
Ethnicity/race composite (number)		–	–
White/Caucasian	84; 37.5%		
Chinese-American	36; 16.1%		
African-American	53; 23.7%		
Hispanic	51; 22.8%		
Other	–		
AHIPSG (events/hour)	28.0±17.6 (range 1 – 88)	13.0±16.7 (range 0 – 113)	50.5±9.1 (range 15 – 115)
Disease severity (number; %)			
None	17; 7.6%	36; 38.8%	0; 0.0%
Mild	17; 7.6%	59; 33.1%	0; 0.0%
Moderate	107; 47.8%	31; 17.4%	4; 16.7%
Severe	83; 37.1%	19; 10.7%	20; 83.3%
Epworth Sleepiness Scale score (/24)	5.8±3.9	14.2±3.7	–
Total sleep time per PSG (hours)	5.9±1.4	5.7±1.1	7.0±1.5

Sleep staging

Cohen's kappa based on an epoch-by-epoch comparison of all sleep stages between Somnolyzer and manual scoring was 0.739 (95% CI 0.737 to 0.741); see Table 2 and Figure 1. The lower-bound of the CI (0.737) exceeded the prespecified threshold of 0.60, supporting this hypothesis. Sleep staging accuracy was 80.7% across all sleep stages (W/N1/N2/N3/R), 94.2% for wake, 87.5% for N1, 86.6% for N2, 95.9% for N3 and 97% for REM. The results of the epoch-by-epoch comparison of sleep stage scoring are presented in Table 3.

Table 2: Comparison of Somnolyzer- and manually-scored sleep staging

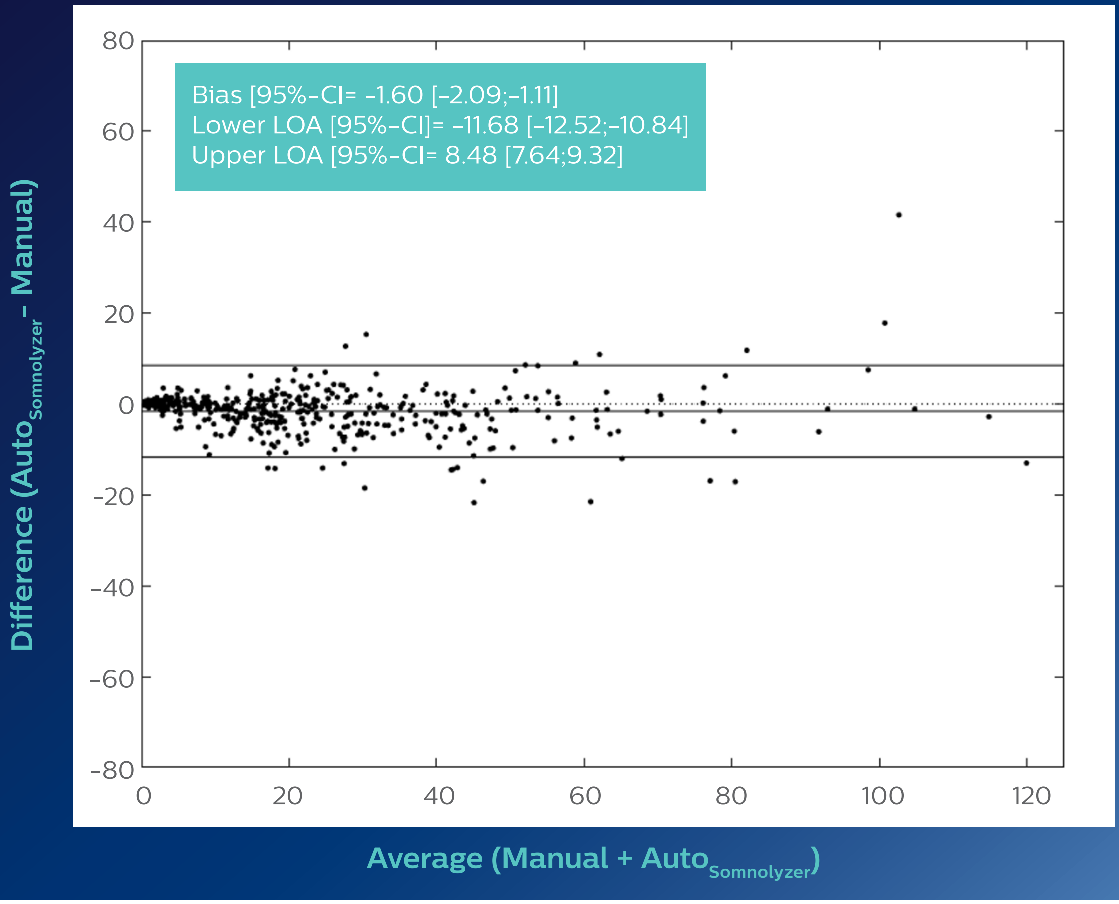
Comparison	Kappa (95% CI)	Accuracy (%)
W/N1/N2/N3/R	0.739 (0.737[a] – 0.741)	80.7
Wake	0.853 (0.851 – 0.855)	94.2
N1	0.457 (0.452 – 0.461)	87.5
N2	0.721 (0.719 – 0.723)	86.8
N3	0.731 (0.727 – 0.735)	95.9
REM	0.868 (0.865 – 0.870)	97.0

[a] The prespecified performance target for this value was 0.6.

The lower limits of each CI can be compared against thresholds defined by Landis and Koch (1977)[15] as follows: 0.0-0.2 slight agreement; 0.21-0.40 fair agreement; 0.41-0.60 moderate agreement; 0.61-0.80 substantial agreement; 0.81-1.0 almost perfect or perfect agreement.

Figure 1: Bland-Altman plot of the Somnolyzer-AHI against the manual-AHI

Mean difference of -1.60 events/hour; lower and upper limits of agreement -11.68 and 8.48, respectively.



Respiratory events

The ICC between the Somnolyzer-AHI and the manual-AHI was 0.969 (95% CI 0.957 to 0.976); see upper part of Table 4. The lower-bound of the CI (0.957) was higher than the prespecified threshold of 0.75, supporting this hypothesis.

Arousals

The ICC between the Somnolyzer-ArI and the manual-ArI was 0.794 (95% CI 0.668 to 0.864); see upper part of Table 4. The lower-bound of the CI (0.668) was higher than the prespecified threshold of 0.50, supporting this hypothesis.

Periodic limb movements

In a sample of 174 PSGs, the ICC between the Somnolyzer-PLMSI and the manual-PLMSI was 0.907 (95% CI 0.877 to 0.930); see upper part of Table 4. The lower-bound of the CI (0.877) was higher than the prespecified threshold of 0.75, supporting this hypothesis.

Table 3: Confusion matrix for epoch-by-epoch sleep staging

	Manual staging				
Somnolyzer staging	Wake	N1	N2	N3	REM
Wake	87.1%	12.2%	1.2%	0.3%	2.6%
N1	9.1%	61.7%	12.3%	0.3%	6.4%
N2	1.9%	22.2%	81.5%	27.2%	4.2%
N3	0.1%	0.1%	4.3%	72.2%	0.0%
REM	1.8%	3.9%	0.8%	0.0%	86.8%

Gray cells indicate the percentage of epochs of each manually-scored sleep stage that was correctly identified by Somnolyzer (that is, sensitivity).

Table 4: Comparison of Somnolyzer- and manually-scored sleep and event metrics

Respiratory, limb movement and desaturation events:

Metric	ICC (95% CI)
AHI (events/hour)	0.969 (0.957[a] – 0.976)
ArI (events/hour)	0.794 (0.668[b] – 0.864)
PLMSI (events/hour)	0.907 (0.877[c] – 0.930)
Total apneas (number)	0.848 (0.733 – 0.904)
Total hypopneas (number)	0.898 (0.757 – 0.946)
ODI (events/hour)	0.990 (0.987 – 0.992)

Sleep staging:

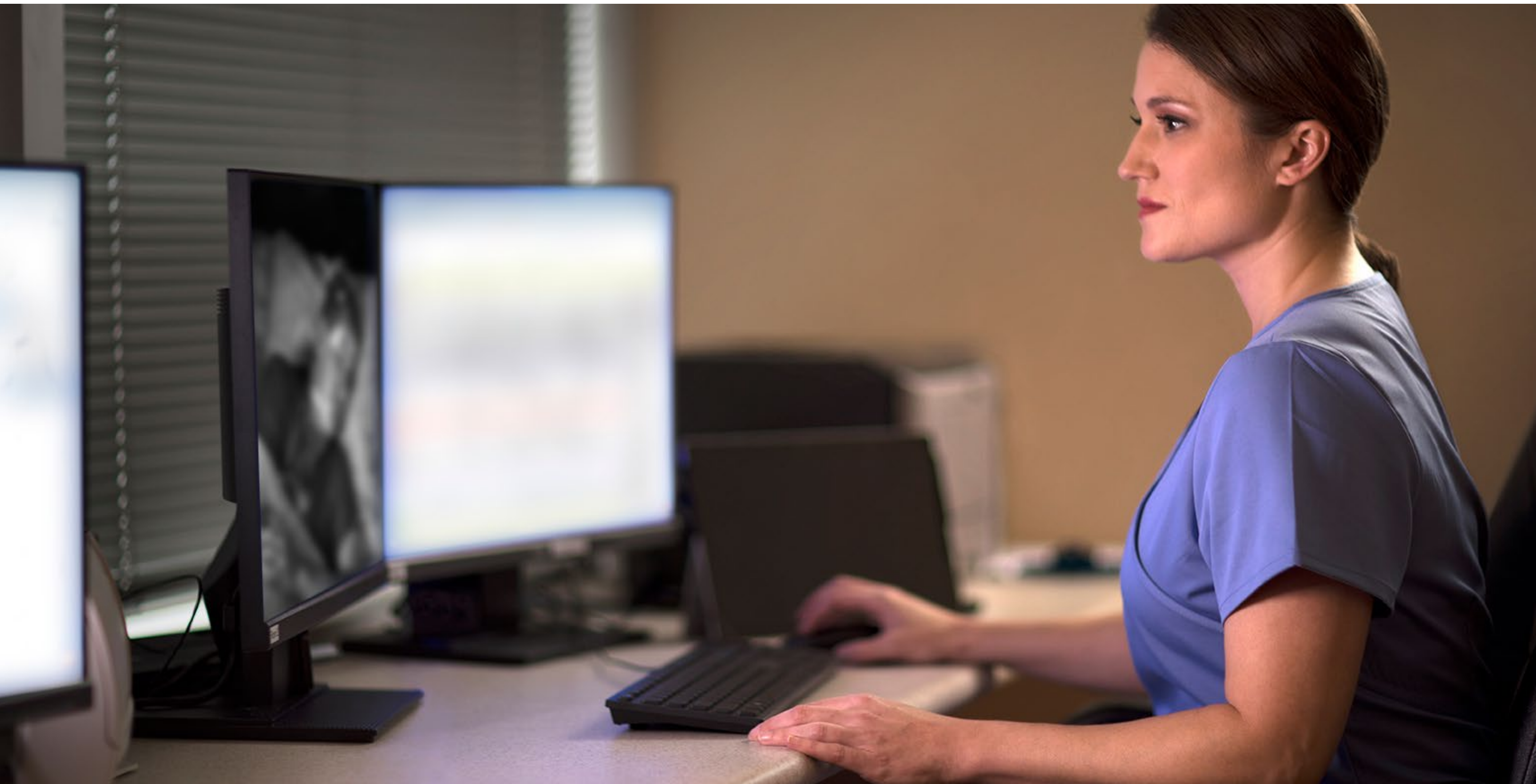
Metric	ICC (95% CI)
Sleep efficiency (%)	0.927 (0.904 – 0.943)
Total sleep time (minutes)	0.938 (0.920 – 0.951)
Time in N1 (minutes)	0.690 (0.372 – 0.826)
Time in N2 (minutes)	0.815 (0.778 – 0.846)
Time in N3 (minutes)	0.772 (0.728 – 0.809)
Time in NREM (minutes)	0.919 (0.875 – 0.944)
Time in REM (minutes)	0.908 (0.888 – 0.925)

[a] The prespecified performance target for this value was 0.75.

[b] The prespecified performance target for this value was 0.5.

[c] The prespecified performance target for this value was 0.75.

The lower limits of each CI can be compared against thresholds defined by Koo and Li (2016),[11] as follows: <0.5 poor reliability; 0.5 to <0.75 moderate reliability, 0.75 to <0.90 good reliability; ≥0.90 excellent reliability.



Discussion

In this study, an automated review of key sleep staging and event detection parameters met prespecified performance thresholds, demonstrating that the Somnolyzer scoring solution generates results that are in agreement with human scoring. Our performance targets were set at the lower margin of the agreement across expert human scorers in the literature. In these studies, Cohen's kappa for discrimination of all sleep stages (W/N1/N2/N3/R), as well as for wake and REM individually, typically show substantial agreement (>0.60), while for N1, N2 and N3, Cohen's kappa values typically show only moderate agreement (>0.4).[12,13] Good inter-rater reliability (ICC >0.75) has been demonstrated for the AHI, TST, sleep efficiency, ODI and PLMSI, while only moderate inter-rater reliability (ICC >0.50) has been demonstrated for the ArI.[5,13,14] By exceeding our performance targets, we can conclude that the variability between Somnolyzer and human scoring in the current study is less than the variability observed across expert scorers and, therefore Somnolyzer scoring provides output that is ready for review and interpretation by a physician. We compare the performance of Somnolyzer 4.1 to that of two recently cleared sleep scoring platforms in Table 5, below. Published results for these two systems included studies with similar methodologies, allowing comparison.

Table 5: Comparison of sleep staging and event detection performance across three auto-scoring systems

Sleep staging sensitivity (% of epochs within each sleep stage correctly identified)

Comparison	Somnolyzer[a]	Nox Sleep System[b]	EnsoSleep[c]
Wake	87%	67%	86%
N1	62%	10%	41%
N2	82%	87%	77%
N3	72%	83%	81%
REM	87%	83%	79%

[a] Results from the current study, copied from Table 3.

[b] Results from the Nox Medical 510(k) K192469; sample size and scoring details unknown. "Nox Sleep System" is Nox Medical's name for the system encompassing Nox A1 Recorder, Nox C1 Access Point and Noxturnal software.

[c] Results from the EnsoData 510(k) K162627; n=72. EnsoSleep is a product of EnsoData, Inc. In this study, median percent agreement was calculated through a bootstrap method in comparison to 2/3 majority sleep staging performed by multiple scorers.

Key strengths of the current study include the large, ethnically/racially diverse sample, which supports the generalizability of algorithm performance, as well as the fact that the PSGs were collected in a range of clinical and research settings using various data collection platforms and montages. Our selection criteria ensured that the performance of Somnolyzer was assessed across the full range of disease severity. The most important limitation of our study was the fact that each PSG was scored by a single technologist. Given the aforementioned variability across scorers,[3,4] our comparator may not represent the true gold standard of manual scoring. Note that, however, numerous different technologists scored the 426 PSGs and thus the aforementioned variability across scorers is reflected in the manual scorings used as the comparator in this study. In that respect, the comparator reflects standard practice in most clinical settings in which a single technologist is responsible for scoring each sleep study.

The Somnolyzer 4.1 automatic scoring solution is a validated tool that offers operational efficiencies and can be easily implemented into existing workflows. The automated scoring results have been shown to be in agreement with human scoring in a variety of settings and with the use of multiple data acquisition technologies. The consistency of an automated scoring process can be beneficial in both clinical and research settings by minimizing inter- and intra-scoring variability. The implementation of Somnolyzer 4.1 has the potential to free up clinical staff to perform other duties and improve the end-to-end sleep care experience; confirmation of these outcomes would be beneficial.

Acknowledgements

We are grateful to the National Sleep Research Resource for making the MESA and SHHS datasets available for these analyses, and acknowledge the National Institutes of Health National Heart Lung and Blood Institute (NHLBI) grant that funds the resource (R24HL114473). MESA was supported by HHSN268201500003I, N01HC95159, N01HC95160, N01HC95161, N01HC95162, N01HC95163, N01HC95164, N01HC95165, N01HC95166, N01HC95167, N01HC95168 and N01HC95169 from the NHLBI, and by cooperative agreements UL1TR000040, UL1TR001079, and UL1TR001420 funded by the National Center for Advancing Translational Sciences. MESA Sleep was supported by R01HL098433 from the NHLBI. SHHS was supported by U01HL53916, U01HL53931, U01HL53934, U01HL53937, U01HL53938, U01HL53940, U01HL53941, U01HL64360 from the NHLBI.

References

1. Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. Lancet Respir Med 2019;7(8):687-698.
2. Hidden health crisis costing America billions: Underdiagnosing and undertreating obstructive sleep apnea draining healthcare system, Online, 2020 AN. Available from the American Academy of Sleep Medicine website: http://www.aasmnet.org/Resources/pdf/sleep-apnea-economic-crisis.pdf. 2016;
3. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 2013;9(1):81-87.
4. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine Inter-scorer Reliability program: respiratory events. J Clin Sleep Med 2014;10(4):447-454.
5. Punjabi NM, Shifa N, Dorffner G, Patil S, Pien G, Aurora RN. Computer-Assisted Automated Scoring of Polysomnograms Using the Somnolyzer System. Sleep 2015;38(10):1555-1566.
6. Zhang GQ, Cui L, Mueller R, et al. The National Sleep Research Resource: towards a sleep data commons. J Am Med Inform Assoc 2018;25(10):1351-1358.
7. Dean DA, 2nd, Goldberger AL, Mueller R, et al. Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. Sleep 2016;39(5):1151-1164.
8. Bakker JP, Tavakkoli A, Rueschman M, et al. Gastric Banding Surgery versus Continuous Positive Airway Pressure for Obstructive Sleep Apnea: A Randomized Controlled Trial. Am J Respir Crit Care Med 2018;197(8):1080-1083.
9. Rosen CL, Auckley D, Benca R, et al. A multisite randomized trial of portable sleep studies and positive airway pressure autotitration versus laboratory-based polysomnography for the diagnosis and treatment of obstructive sleep apnea: the HomePAP study. Sleep 2012;35(6):757-767.
10. Chen X, Wang R, Zee P, et al. Racial/Ethnic Differences in Sleep Disturbances: The Multi-Ethnic Study of Atherosclerosis (MESA). Sleep 2015;38(6):877-888.
11. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15(2):155-163.
12. Danker-Hopfe H, Anderer P, Zeitlhofer J, et al. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. J Sleep Res 2009;18(1):74-84.
13. Magalang UJ, Chen NH, Cistulli PA, et al. Agreement in the scoring of respiratory events and sleep among international sleep centers. Sleep 2013;36(4):591-596.
14. Malhotra A, Younes M, Kuna ST, et al. Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. Sleep 2013;36(4):573-582.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33(1):159-174.

www.philips.com/respironics

Caution: U.S. federal law restricts these devices to sale by or on the order of a physician.

edoc BG 5/20/21 MC 4110693 v01
6501 Living Place, Pittsburgh, PA 15206
800 345 6443 • 724 387 4000