JCSM | Journal *of* Clinical Sleep Medicine

SCIENTIFIC INVESTIGATIONS

Correlations between home sleep apnea tests and polysomnography outcomes do not fully reflect the diagnostic accuracy of these tests

Frederik Massie, MSc; Bart Van Pee, MSc; Jeroen Bergmann, MSc, PhD

Natural Interaction Lab, Thom Building, Department of Engineering, University of Oxford, Oxford, United Kingdom

Study Objectives: The clinical performance of home sleep apnea tests (HSATs) can be described by their (diagnostic) accuracy, defined as the percentage agreement with the obstructive sleep apnea severity category (normal, mild, moderate, and severe) based on polysomnography. Rather than reporting on accuracy, there has been a strong reliance in the literature to report correlation coefficients between the apnea-hypopnea index of HSATs and polysomnography to support claims of diagnostic performance. This is surprising, as it has been well described that correlation coefficients are inadequate to evaluate equivalence between 2 parameters. The aim of this study was to systematically investigate the magnitude of the discrepancies between correlation coefficients and diagnostic accuracy reported in or retrievable from HSAT validation studies.
Methods: We compared the discrepancy between accuracy and apnea-hypopnea index correlation coefficients of all validation papers that met the inclusion criteria. A total of 20 papers were retained, representing a participant pool of 1,652.
Results: The weighted average apnea-hypopnea index correlation across all 20 papers was 0.82 and the weighted average accuracy was 0.61, highlighting a discrepancy of 0.21 and an overall misdiagnosis rate of 39%.
Conclusions: The results of our study confirm the need for increased scientific rigor in selecting primary performance endpoints to support clinical performance claims of HSATs.
Keywords: meta-analysis, accuracy, correlation, diagnosis, home sleep apnea testing
Citation: Massie F, Van Pee B, Bergmann J. Correlations between home sleep apnea tests and polysomnography outcomes do not fully reflect the diagnostic accuracy of these tests. *J Clin Sleep Med*. 2022;18(3):871–876.

BRIEF SUMMARY
Current Knowledge/Study Rationale: Rather than reporting on diagnostic accuracy, there has been a strong reliance in the home sleep apnea testing literature on correlation coefficients between the apnea-hypopnea index of home sleep apnea tests and polysomnography as primary endpoint parameters to support claims of diagnostic performance. Nevertheless, it is known that correlation coefficients are inadequate to evaluate equivalence between 2 parameters. It was the aim of our study to systematically investigate the magnitude of the discrepancies between correlation coefficients and diagnostic accuracy reported in or retrievable from home sleep apnea test validation studies.
Study Impact: Our meta-analysis revealed a discrepancy of 21% between the apnea-hypopnea index correlation and diagnostic accuracy. This highlights the need for increased scientific rigor in selecting performance endpoints to describe the performance of home sleep apnea tests.

INTRODUCTION

The American Academy of Sleep Medicine (AASM) clinical guideline for the evaluation, management, and long-term care of obstructive sleep apnea (OSA) in adults[1] defines the severity of OSA by whether the apnea-hypopnea index (AHI) is lower than 5 (normal), between 5 and 15 (mild OSA), between 15 and 30 (moderate OSA), or above 30 events per hour of sleep (severe OSA).

The performance of a home sleep apnea test (HSAT) can be evaluated by determining the percentage agreement of the OSA severity (normal, mild, moderate, or severe OSA) estimated by the HSAT with that determined by concurrently administered gold-standard polysomnography (PSG). This percentage agreement is commonly referred to as the (diagnostic) accuracy or the (diagnostic) concordance.

Yalamanchali et al[2] performed a comprehensive meta-analysis of the clinical performance of the most frequently deployed

HSAT by systematically reviewing all papers that reported on the AHI correlation between the HSAT and PSG. From the apparently very strong correlation coefficients reported in the included papers, the authors concluded that the HSAT represents a viable alternative to PSG for confirmation of clinically suspected sleep apnea.

However, it has been well described in domains outside of sleep medicine that correlation coefficients such as Pearson's and Spearman's correlations are inadequate in describing the degree of one-on-one correspondence of 2 diagnostic endpoint parameters.[3] Almost 2 decades ago, Flemons et al[4,5] pointed out, based on an example comparing the respiratory disturbance index (RDI) of HSAT and the AHI of PSG, how (Pearson) correlation coefficients are only able to indicate whether 2 measurements are (linearly) related, and not whether they have a similar magnitude.

Indeed, the Pearson correlation coefficient attains the maximum value of 1 upon a perfect linear relationship between 2

endpoint parameters, but it does not penalize a constant offset or scaling factor between the parameters. Worse in this context is the Spearman correlation coefficient, as it attains the maximum value of 1 when there is a perfect monotonously increasing relationship between the 2 parameters, without penalizing for the nonlinearity of such a relationship. Correlation coefficients are also heavily influenced by extreme datapoints, such as very high AHIs.

These issues can be illustrated by the examples presented in **Figure 1**. All 4 panels show how AHI correlation coefficients can be arbitrarily high for devices with little to no clinical utility (as illustrated by diagnostic accuracies of < 0.6). The top left and right panels, respectively, show how the Pearson correlation does not take scaling mismatches and constant offsets between the AHIs of HSAT and PSG into account. The bottom left panel shows how low AHIs (< 5 events/h) and extreme AHIs (> 30 events/h) skew Pearson correlations to misleadingly large values, while the intrinsic diagnostic accuracy is very low. The bottom left panel has the same diagnostic accuracy of 0.5 as the bottom right panel but has fewer extreme AHI values in the sample pool, highlighting the collapse of the Pearson correlation in the absence of extremity-introduced bias. Importantly, extremely low and high AHIs, while excessively influencing correlations, have the least relevance in the performance evaluation of an HSAT. Indeed, it is comparatively straightforward to detect a near-complete absence of respiratory events (eg, by a completely flat oxygen saturation [SpO_2] trace) or an extreme prevalence of respiratory events (eg, by the presence of extremely frequent oxygen desaturations). The patients most challenging to assess with any HSAT device are those with AHIs between 5 and 30 events/h, a range of particular importance for therapeutic decisions.

A recent study by Ioachimescu et al,[6] in which the HSAT device described by Yalamanchali et al[2] was evaluated against

Figure 1—Sample scatterplots highlighting issues of relying on correlation coefficients to describe equivalence between 2 parameters.
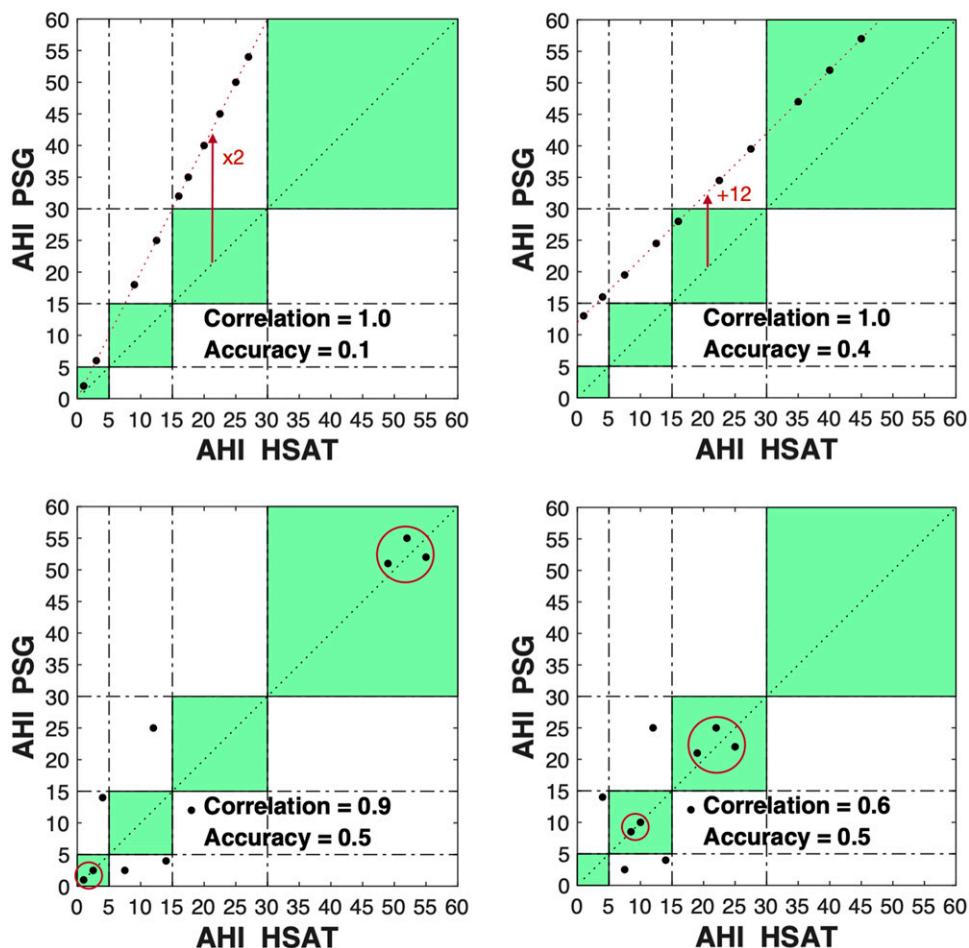
The green-shaded squares mark the areas in which PSG and HSAT agree on the AHI severity category. Top left: AHIs differ by a constant scaling factor (AHI PSG = 2 × AHI HSAT). Top right: AHIs differ by a constant offset (AHI PSG = AHI HSAT + 12). The differences between AHI PSG and AHI HSAT do not impact the correlation coefficient but strongly impact the accuracy. Bottom: Extreme values disproportionally influence correlation coefficients. Moving extreme values (dots within red circles) more to the center of the spectrum, without altering the accuracy, has a strong effect on the correlation coefficient. This is observed by comparing the bottom left (correlation = 0.9) and bottom right (correlation = 0.6) panels. AHI = apnea-hypopnea index, HSAT = home sleep apnea test, PSG = polysomnography.

PSG on 500 participants, reported a very apparent discrepancy between the study's AHI Pearson correlation of 0.80 and diagnostic accuracy of 0.53. These findings drove a starkly different conclusion than the one by Yalamanchali et al.[2] Indeed, the authors stated that because of the large diagnostic misclassification rate, patients without OSA or mild disease assessed by the HSAT should undergo repeat in-laboratory PSG as such, limiting the device's utility to inclusion screening. These conflicting conclusions highlight the peril of reliance on correlation coefficients as opposed to diagnostic accuracy as a primary clinical performance endpoint of HSAT devices.

Prompted by the scientific community's surprise[7] about these findings, we systematically compared the discrepancy between diagnostic accuracy and AHI correlation coefficients of all published validation studies of the HSAT devices that met the search criteria.

METHODS

In order to allow for cross-trial, weighted-average aggregation of accuracy and correlation parameters, we limited the inclusion of HSAT types to the single most frequently validated HSAT device. The WatchPAT device (Itamar Medical, Caesarea, Israel) was found to be the most frequently included in clinical validation studies (∼ 20% of all HSAT validation studies report on this device).

As such, we searched PubMed on April 3, 2021, for the search term "'watchpat' or 'watch-pat' or 'watch pat.'"

Any paper that met at least 1 of the exclusion criteria was discarded from further analysis. The following exclusion criteria were used:

1. The paper was not available in the English language.
2. The paper did not report on the performance validation of the HSAT device.
3. The HSAT was not concurrently administered with the PSG.
4. No diagnostic accuracy could be retrieved from the paper.
5. No AHI correlation coefficient could be retrieved from the paper.

For all included papers, the correlation between the (automated) AHI estimate of HSAT and PSG was retrieved. We additionally registered the type of the reported correlation coefficient (Spearman, Pearson, or intraclass correlation coefficient). For all included papers, we retrieved the accuracy of the HSAT. For papers for which the accuracy was not explicitly reported, or for which it could not be precisely calculated from the confusion matrix, we estimated the accuracy based on the datapoints of the scatter plots or the Bland-Altman plots, or inverse-calculated the accuracy from the disease prevalence, sensitivity, and specificity values. Such scatterplots are equivalent to those presented in **Figure 1**. We counted the number of dots falling in the green rectangles and divided this number by the total number of dots to arrive at the accuracy value. Any discrepancy between the estimated participant size from the scatter plots or Bland-Altman plots and the reported participant size was also identified.

As accuracy estimates could be impacted by incomplete or incorrect visual retrieval of the datapoints, we performed a subgroup analysis in which papers that did not allow exact diagnostic accuracy retrieval were excluded.

For papers in which only an RDI correlation and accuracy could be retrieved, we used the RDI instead of the AHI. Finally, the scoring rules used to score the AHI by PSG were noted for each paper. We performed a subgroup analysis on papers that reported on the current (2012) AASM respiratory event scoring rules.[8]

We charted out the diagnostic accuracy and the AHI correlation coefficient for each included paper, ranked by decreasing study participant number. Finally, we calculated the overall (sample-size weighted average) accuracy and the AHI correlation coefficient of the total participant pool.

RESULTS

As illustrated in **Figure 2**, the search term resulted in the identification of 86 papers, of which 20 were retained after application of the exclusion criteria. Fifty out of 86 identified papers were not considered HSAT validation studies, rendering this the main reason for exclusion. The results of the systematic analysis are summarized in **Figure 3** and **Table 1**.

The weighted average AHI correlation across all 20 papers was 0.82 and the weighted average accuracy was 0.61, highlighting a discrepancy of 0.21. The lowest observed accuracy was 0.35 (correlation 0.76) and the highest observed accuracy was 0.95 (correlation 0.95), which illustrates a large degree of performance variation between studies. The lowest observed AHI correlation was 0.65 (accuracy 0.61) and the highest observed AHI correlation was 0.96 (accuracy 0.74).

The weighted average AHI correlation and accuracy across all papers for which the accuracy could exactly be retrieved were, respectively, 0.82 and 0.58. For 6 out of 20 papers, the accuracy was exactly retrievable. Six out of 20 studies reported on the AASM 2020 scoring rules using the 1A rule for hypopnea

Figure 2—Overview of identified and retained papers and the reasons for any exclusions.
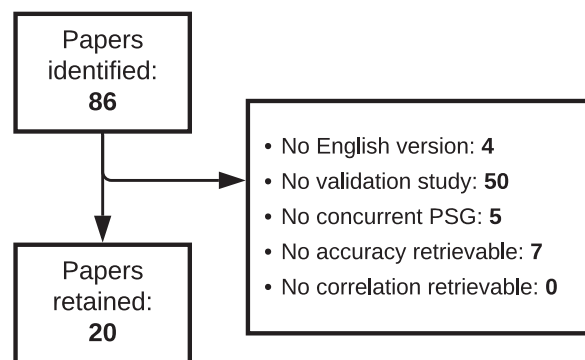
PSG = polysomnography.

Figure 3—Comparison of accuracy and correlations for each included paper.
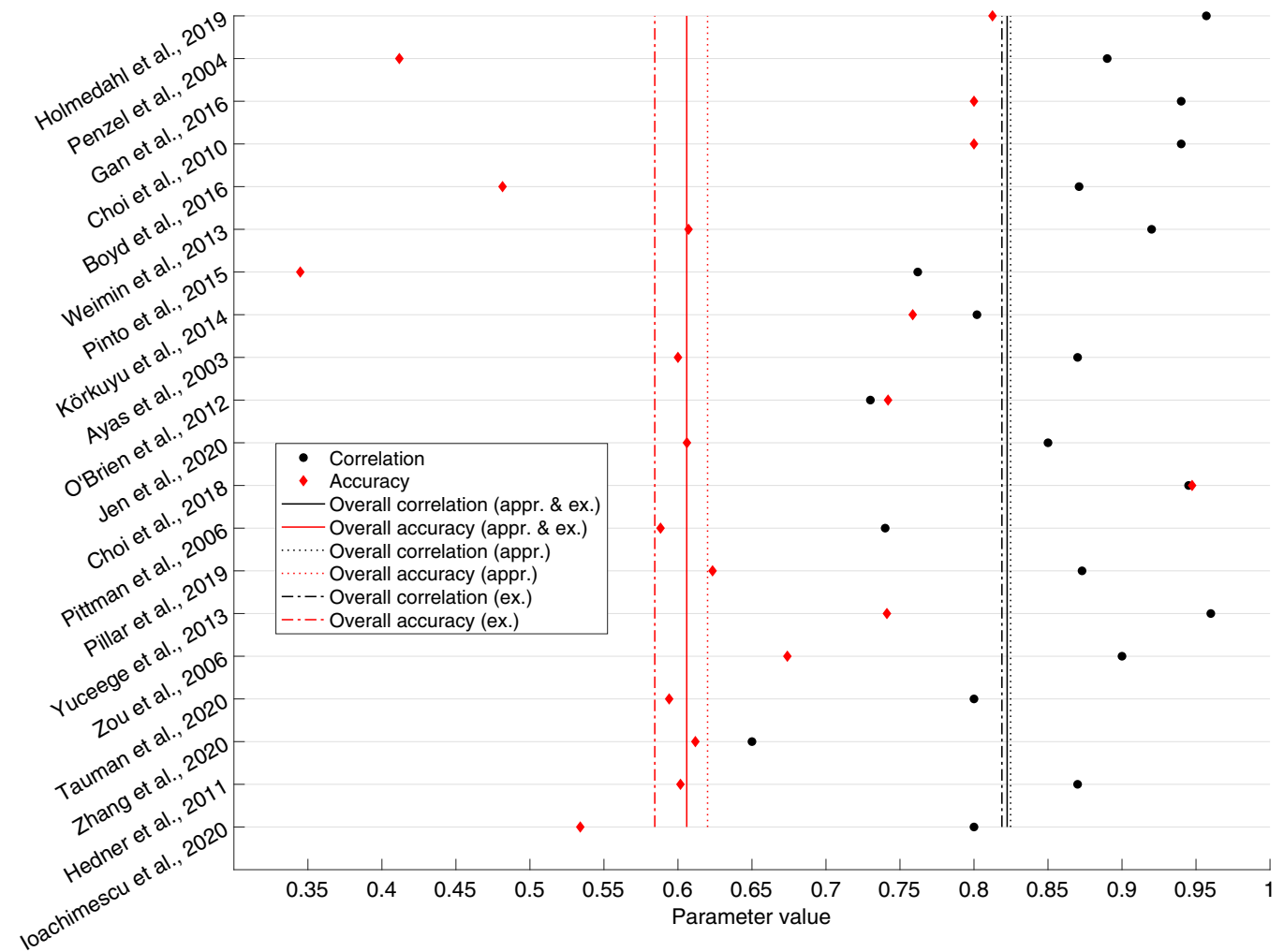
Comparison of accuracy and AHI correlations (horizontal axis) as retrieved from each included paper (vertical axis). Papers are ranked by increasing sample size (largest at bottom). Overall accuracy and correlation were calculated from a weighted average of all individual trials, with the weights equal to the trial population size. AHI = apnea-hypopnea index, appr. = accuracy was approximated from the graphs or inverse-calculated, ex. = accuracy was exactly inferred.

requiring 3% desaturations.[8] Only 1 study reported (partially) on the AASM 2020 scoring rules using the 1B rule for hypopnea requiring 4% desaturations. Comparing the studies that scored using any of the AASM 2012 rules with the papers that applied an older scoring rule revealed a difference of 0.04 (from 0.59 to 0.63) for accuracy and 0.08 (0.79 to 0.87), for correlation, where the highest number was found for the older rules.

DISCUSSION

Similar to the results reported by Ioachimescu et al,[6] our meta-analysis determined a discrepancy of 21% between the diagnostic accuracy of 61% and correlation coefficient of 82%, derived from a total participant pool size of 1,652. When only considering papers reporting on the latest AASM scoring rules, a misdiagnosis rate of 41% was found. For only 6 out of the 20 included papers, the diagnostic accuracy could directly be

retrieved or exactly calculated, revealing a large underreporting of this important endpoint parameter.

These observations confirm the urgent need for researchers and device manufacturers to increase scientific rigor and transparency in the presentation of HSAT devices' clinical performance by complementing their reporting with diagnostic accuracy, in a move away from the overreliance on correlations. This study can be viewed as a precursor to much-needed future work on the determination of a statistically robust set of parameters to characterize the performance of HSATs in general, and the agreement between HSATs and PSG in particular. This problem is embedded within the broader observation that hypothesis testing is typically being deployed for the determination of significant differences between 2 measurements, and not for the determination of their equivalence.

This meta-analysis had several limitations. First, the accuracy had to be estimated from alternative data such as scatter plots and Bland-Altman plots for 14 out of 20 papers, due to

Table 1—Tabulation of accuracy and correlations for each included paper.

Study	AHI or RDI	Correlation (type)	Accuracy (appr. or exp.)	Difference	n (reported)	Scoring Rules
Holmedahl et al, 2019[9]	AHI	0.96 (ICC)	0.81 (appr.)	0.15	16	O
Penzel et al, 2004[10]	RDI	0.89 (Pearson)	0.41 (appr.)	0.48	17	O
Gan et al, 2016[11]	AHI	0.94 (Pearson)	0.80 (appr.)	0.14	20	A
Choi et al, 2010[12]	AHI	0.94 (Pearson)	0.80 (ex.)	0.14	25	O
Boyd et al, 2016[13]	AHI	0.87 (Pearson)	0.48 (appr.)	0.39	27 (28)	O
Weimin et al, 2013[14]	AHI	0.92 (Pearson)	0.61 (ex.)	0.31	28	O
Pinto et al, 2015[15]	AHI	0.76 (Spearman)	0.35 (appr.)	0.42	29 (30)	O
Körkuyu et al, 2014[16]	AHI	0.80 (Spearman)	0.76 (appr.)	0.04	29 (30)	O
Ayas et al, 2003[17]	AHI	0.87 (Pearson)	0.60 (ex.)	0.27	30	O
O'Brien et al, 2012[18]	AHI	0.73 (Pearson)	0.74 (ex.)	−0.01	31	O
Jen et al, 2020[19]	AHI	0.85 (Pearson)	0.61 (appr.)	0.24	33	O
Choi et al, 2018[20]	AHI	0.95 (Spearman)	0.95 (ex.)	0.00	38	A
Pittman et al, 2006[21]	AHI	0.74 (ICC)	0.59 (appr.)	0.15	68 (70)	O
Pillar et al, 2019[22]	AHI	0.87 (Pearson)	0.62 (appr.)	0.25	77 (84)	A
Yuceege et al, 2013[23]	AHI	0.96 (Pearson)	0.74 (appr.)	0.22	85	O
Zou et al, 2006[24]	AHI	0.90 (Pearson)	0.67 (appr.)	0.23	92	O
Tauman et al, 2020[25]	AHI	0.80 (Pearson)	0.59 (appr.)	0.21	101	A
Zhang et al, 2020[26]	AHI	0.65 (Spearman)	0.61 (appr.)	0.04	170	A
Hedner et al, 2011[27]	RDI	0.87 (ICC)	0.60 (appr.)	0.27	236	O
Ioachimescu et al, 2020[6]	AHI	0.80 (Pearson)	0.53 (ex.)	0.27	500	A and B
Overall (A or B)	AHI	0.79	0.59	0.20	906	A or B
Overall (O)	AHI or RDI	0.87	0.63	0.24	746	O
Overall (appr.)	AHI or RDI	0.83	0.62 (appr.)	0.21	1,000	A, B, or O
Overall (ex.)	AHI or RDI	0.82	0.58 (ex.)	0.24	652	A, B, or O
Overall	AHI or RDI	0.82	0.61	0.21	1,652	A, B, or O

Papers are ranked by increasing sample size. Papers exclusively reporting on the RDI were highlighted as such in column 2. For "appr." papers, the participant number (n) is displayed based on the paper's graphs, with the reported number between parentheses when it was different to the sample size determined from the graphs or metrics used to estimate the accuracy. "Overall (A or B)" and "Overall (O)," respectively, contain the sample-size weighted average of all studies scored by "A or B" rules or "O" rules. "Overall (appr.)" and "Overall (ex.)," respectively, contain the sample-size weighted average of all studies with an "appr." or "ex." accuracy. A = AASM 2020 scoring rules with 1A rule for hypopnea, AHI = apnea-hypopnea index, appr. = accuracy was approximated from the graphs or inverse-calculated, B = AASM 2020 scoring rules with 1B rule for hypopnea, ex. = accuracy was exactly inferred, ICC = intraclass correlation coefficient, O = scoring rule other than any AASM 2012 rule, RDI = respiratory disturbance index.

lack of explicit reporting of the parameter. Second, only 6 papers reported on the latest 2012 AASM scoring rules and only 1 paper reported on the latest 1B rules for the scoring of hypopnea. Finally, there was heterogeneity in participant population characteristics. Most papers (11/20) reported on a participant population with a high suspicion of OSA, while other papers discussed specific participant populations, such as those with cardiovascular or pulmonary comorbidities. Nevertheless, the totality of participants represents a patient group that closely aligns with the HSAT's utilization in the market.

ABBREVIATIONS

AASM, American Academy of Sleep Medicine
AHI, apnea-hypopnea index
HSAT, home sleep apnea test

OSA, obstructive sleep apnea
PSG, polysomnography
RDI, respiratory disturbance index

REFERENCES

1. Adult Obstructive Sleep Apnea Task Force of the American Academy of Sleep Medicine. Clinical guideline for the evaluation, management and long-term care of obstructive sleep apnea in adults. *J Clin Sleep Med.* 2009;5(3):263–276.

2. 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.

3. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. *Anesth Analg.* 2018;126(5):1763–1768.

4. Flemons WW, Littner MR, Rowley JA, et al. Home diagnosis of sleep apnea: a systematic review of the literature. An evidence review cosponsored by the American Academy of Sleep Medicine, the American College of Chest Physicians, and the American Thoracic Society. *Chest.* 2003;124(4):1543–1579.

5. Flemons WW, Littner MR. Measuring agreement between diagnostic devices. *Chest.* 2003;124(4):1535–1542.

6. Ioachimescu OC, Allam JS, Samarghandi A, et al. Performance of peripheral arterial tonometry-based testing for the diagnosis of obstructive sleep apnea in a large sleep clinic cohort. *J Clin Sleep Med.* 2020;16(10):1663–1674.

7. Kirsch D. Autopilot and algorithms: accidents, errors, and the current need for human oversight. *J Clin Sleep Med.* 2020;16(10):1651–1652.

8. Berry RB, Quan SF, Abreu A, 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 2.6. Darien, IL: American Academy of Sleep Medicine; 2020.

9. Holmedahl NH, Fjeldstad OM, Engan H, Saxvig IW, Grønli J. Validation of peripheral arterial tonometry as tool for sleep assessment in chronic obstructive pulmonary disease. *Sci Rep.* 2019;9(1):19392.

10. Penzel T, Kesper K, Pinnow I, Becker HF, Vogelmeier C. Peripheral arterial tonometry, oximetry and actigraphy for ambulatory recording of sleep apnea. *Physiol Meas.* 2004;25(4):1025–1036.

11. 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.

12. Choi JH, Kim EJ, Kim YS, et al. Validation study of portable device for the diagnosis of obstructive sleep apnea according to the new AASM scoring criteria: Watch-PAT 100. *Acta Otolaryngol.* 2010;130(7):838–843.

13. Boyd SB, Upender R, Walters AS, et al. Effective apnea-hypopnea index ("effective AHI"): a new measure of effectiveness for positive airway pressure therapy. *Sleep.* 2016;39(11):1961–1972.

14. Weimin L, Rongguang W, Dongyan H, Xiaoli L, Wei J, Shiming Y. Assessment of a portable monitoring device WatchPAT 200 in the diagnosis of obstructive sleep apnea. *Eur Arch Otorhinolaryngol.* 2013;270(12):3099–3105.

15. Pinto JA, Godoy LB, Ribeiro RC, Mizoguchi EI, Hirsch LAM, Gomes LM. Accuracy of peripheral arterial tonometry in the diagnosis of obstructive sleep apnea. *Braz J Otorhinolaryngol.* 2015;81(5):473–478.

16. Körkuyu E, Düzlü M, Karamert R, et al. The efficacy of Watch PAT in obstructive sleep apnea syndrome diagnosis. *Eur Arch Otorhinolaryngol.* 2015; 272(1):111–116.

17. Ayas NT, Pittman S, MacDonald M, White DP. Assessment of a wrist-worn device in the detection of obstructive sleep apnea. *Sleep Med.* 2003;4(5):435–442.

18. 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.

19. Jen R, Orr JE, Li Y, et al. Accuracy of WatchPAT for the diagnosis of obstructive sleep apnea in patients with chronic obstructive pulmonary disease. *COPD.* 2020; 17(1):34–39.

20. Choi JH, Lee B, Lee JY, Kim HJ. Validating the Watch-PAT for diagnosing obstructive sleep apnea in adolescents. *J Clin Sleep Med.* 2018;14(10):1741–1747.

21. Pittman SD, Pillar G, Berry RB, Malhotra A, MacDonald MM, White DP. Follow-up assessment of CPAP efficacy in patients with obstructive sleep apnea using an ambulatory device based on peripheral arterial tonometry. *Sleep Breath.* 2006; 10(3):123–131.

22. Pillar G, Berall M, Berry R, et al. Detecting central sleep apnea in adult patients using WatchPAT—a multicenter validation study. *Sleep Breath.* 2020;24(1): 387–398.

23. Yuceege M, Firat H, Demir A, Ardic S. Reliability of the Watch-PAT 200 in detecting sleep apnea in highway bus drivers. *J Clin Sleep Med.* 2013;9(4):339–344.

24. 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.

25. Tauman R, Berall M, Berry R, et al. Watch-PAT is useful in the diagnosis of sleep apnea in patients with atrial fibrillation. *Nat Sci Sleep.* 2020;12: 1115–1121.

26. 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.

27. 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.

SUBMISSION & CORRESPONDENCE INFORMATION

DISCLOSURE STATEMENT