HHS Public Access

Author manuscript

Sleep Breath. Author manuscript; available in PMC 2019 July 09.

Published in final edited form as:

Sleep Breath. 2019 March ; 23(1): 25–31. doi:10.1007/s11325-018-1715-6.

Home sleep apnea testing: comparison of manual and automated scoring across international sleep centers

Ulysses J. Magalang[1,2], **Jennica N. Johns**[1], **Katherine A. Wood**[1], **Jesse W. Mindel**[1], **Diane C. Lim**[3], **Lia R. Bittencourt**[4], **Ning-Hung Chen**[5], **Peter A. Cistulli**[6,7], **Thorarinn Gíslason**[8,9], **Erna S. Arnardottir**[8,9], **Thomas Penzel**[10], **Sergio Tufik**[4], and **Allan I. Pack**[3]

[1]Division of Pulmonary, Allergy, Critical Care, and Sleep Medicine, The Ohio State University Wexner Medical Center, 201 Davis Heart and Lung Research Institute, 473 West 12th Avenue, Columbus, OH 43210, USA [2]Neuroscience Research Institute, The Ohio State University Wexner Medical Center, Columbus, OH, USA [3]Center for Sleep and Circadian Neurobiology, Division of Sleep Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA [4]Departamento de Psicobiologia, Universidade Federal de São Paulo, São Paulo, Brazil [5]Division of Pulmonary, Critical Care, and Sleep Medicine, Chang Gung Memorial Hospital, Taoyuan, Taiwan [6]Charles Perkins Centre, University of Sydney, Camperdown, Australia [7]Department of Respiratory and Sleep Medicine, Royal North Shore Hospital, Sydney, Australia [8]Department of Sleep Medicine, Landspitali University Hospital, Reykjavik, Iceland [9]Medical Faculty, University of Iceland, Reykjavik, Iceland [10]Interdisciplinary Center of Sleep Medicine, Charité University Hospital, Berlin, Germany

Abstract

Purpose—To determine the agreement between the manual scoring of home sleep apnea tests (HSATs) by international sleep technologists and automated scoring systems.

Methods—Fifteen HSATs, previously recorded using a type 3 monitor, were saved in European Data Format. The studies were scored by nine experienced technologists from the sleep centers of the Sleep Apnea Global Interdisciplinary Consortium (SAGIC) using the locally available software. Each study was scored separately by human scorers using the nasal pressure (NP), flow derived from the NP signal (transformed NP), or respiratory inductive plethysmography (RIP) flow. The same procedure was followed using two automated scoring systems: Remlogic (RLG) and Noxturnal (NOX).

Ulysses J. Magalang ulysses.magalang@osumc.edu.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Formal consent is not required for this type of study.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11325-018-1715-6) contains supplementary material, which is available to authorized users.

Results—The intra-class correlation coefficients (ICCs) of the apnea-hypopnea index (AHI) scoring using the NP, transformed NP, and RIP flow were 0.96 [95% CI 0.93–0.99], 0.98 [0.96–0.99], and 0.97 [0.95–0.99], respectively. Using the NP signal, the mean differences in AHI between the average of the manual scoring and the automated systems were – 0.9 ± 3.1/h (AHI_{RLG} vs AHI_{MANUAL}) and – 1.3 ± 2.6/h (AHI_{NOX} vs AHI_{MANUAL}). Using the transformed NP, the mean differences in AHI were – 1.9 ± 3.3/h (AHI_{RLG} vs AHI_{MANUAL}) and 1.6 ± 3.0/h (AHI_{NOX} vs AHI_{MANUAL}). Using the RIP flow, the mean differences in AHI were – 2.7 ± 4.5/h (AHI_{RLG} vs AHI_{MANUAL}) and 2.3 ± 3.4/h (AHI_{NOX} VS AHI_{MANUAL}).

Conclusions—There is very strong agreement in the scoring of the AHI for HSATs between the automated systems and experienced international technologists. Automated scoring of HSATs using commercially available software may be useful to standardize scoring in future endeavors involving international sleep centers.

Keywords

Sleep apnea; Automation; Computer-assisted diagnosis

Introduction

Home sleep apnea testing (HSAT) is now commonly used around the world to diagnose obstructive sleep apnea (OSA) compared to in-laboratory polysomnography (PSG) [1–4]. HSAT is performed most often using a type 3 portable device that monitors respiration during the normal sleep hours. The home sleep study is unattended and does not record sleep stages [5]. The signals used as a measure of airflow in HSATs are the nasal pressure (NP) signal, the square root transformation of the nasal pressure signal (transformed NP), or alternatively, the uncalibrated respiratory inductive plethysmography (RIP) flow which is recommended as a surrogate measure if the NP signal is inadequate [6]. Our group previously showed that there was a substantial agreement in the scoring of the respiratory events for HSAT among international sleep technologists in the Sleep Apnea Global Interdisciplinary Consortium (SAGIC) using any of these signals as a measure of airflow [7].

Automated scoring systems of respiratory events are included in commercially available software and may save time as well as standardize the scoring of HSATs due to the use of computerized algorithms. Compared to manual scoring, two prior studies indicate that automated scoring underestimates the apnea-hypopnea index (AHI) with a mean difference ranging from 5 to 8 events/h [8, 9]. However, both these studies compared the automated systems to only one human scorer. Thus, these studies failed to take into consideration the scoring variability of respiratory events among human scorers. A more appropriate comparison of the real-world performance of automated systems should include a group of human scorers to account for the known variation in human scoring. In addition, the agreement between manual scoring of HSATs by international sleep technologists and automated scoring systems has not been previously reported. This would be important to determine given the increasing clinical use of HSAT and the developing collaboration among sleep researchers worldwide in multi-center international studies [10, 11].

We aimed to extend the findings of our previously published study on the HSAT scoring agreement among international sleep technologists [7], and examined the agreement between manual and automated scoring systems using different types of airflow signals. Our primary hypothesis was that there would be strong agreement between the automated systems and manual scoring.

Methods

The SAGIC centers that participated in this study are Sydney, Australia; Sao Paulo, Brazil; Berlin, Germany; Reykjavik, Iceland; Taipei, Taiwan; Columbus, OH, USA; and Philadelphia, PA, USA. Approval for the study was provided by the Institutional Review Board of the Ohio State University Wexner Medical Center. A waiver of informed consent was obtained because the HSATs were previously recorded and de-identified.

Home sleep studies

Fifteen HSATs that were recorded in Columbus, OH, were chosen. Previously scored HSATs are routinely stored at the Ohio State University Sleep Disorders Center on a quarterly basis in a secure electronic drive. Studies were randomly selected from one folder containing the list of patients who had HSATs obtained in one quarter (74 studies). Fifteen studies were chosen according to a power analysis as explained below. Exclusion criteria included studies performed while on positive airway pressure therapy or while wearing a mandibular advancement device, or if the clinical report stated that the study was not interpretable because of excessive artifacts or absence of adequate data. During the process of choosing the 15 studies, three studies were excluded because they were performed while on CPAP therapy and two were excluded because they were performed while wearing a dental device. All HSATs were judged to have adequate signal quality. After the 15th HSAT was chosen, no other studies were evaluated for inclusion [7].

All HSATs were originally recorded using an Embletta Gold type 3 portable device (Natus Neurology, Tonawanda, NY). All patients were seen by a technologist, during a scheduled session, at which time-detailed instructions and demonstration of how the sensors should be applied were performed. In addition, patients were asked to demonstrate to the technologist how they will use the HSAT equipment on themselves. The patient was also provided with written instructions to take home at the end of the session. Signals were recorded using the following sampling rates: nasal pressure, 50 Hz; chest and abdominal movement, 50 Hz; and pulse oximetry, 3 Hz. Pulse oximetry (Nonin Medical, Inc., Plymouth, MN) used an averaging algorithm of 3 s or faster for pulse rates of 60 beats per minute or greater. The HSATs were converted into European Data Format (EDF) [12]. All prior scoring of the HSATs were removed with the EDF conversion.

Manual scoring

The HSAT EDF files were imported into the local software used for scoring at each site. The scoring software used for manual scoring were Remlogic (Natus Neurology, Tonawanda, NY) [Berlin, Sao Paulo, Philadelphia, and Columbus sites], Compumedics (Compumedics Limited, Victoria, Australia) [Sydney and Taipei sites], and Noxturnal (Nox Medical,

Reykjavik, Iceland) [Reykjavik site]. The guidelines for scoring were summarized in Microsoft PowerPoint and Microsoft Word formats and provided to the nine human scorers. This was supplemented by an educational online conference (WebEx, Cisco Systems, Inc., San Jose, CA) headed by an investigator (UJM) involving the scorers. All nine scorers had at least 5 years of experience in scoring clinical sleep studies and were chosen by the investigator at each participating SAGIC site (two scorers in Berlin, two scorers in São Paulo, and one scorer each in Philadelphia, Columbus, Sydney, Taipei, and Reykjavik). The analysis start time ("lights out") and analysis end time ("lights on") were provided for each HSAT. Respiratory events were manually scored for each 30-s epoch according to standard procedures [7]. An apnea was defined as a decrease in airflow sensor excursion by ≥ 90% of baseline lasting at least 10 s with ≥ 90% of the event duration meeting the amplitude reduction criteria. An apnea was considered to be obstructive if there was visible respiratory effort throughout the entire period of absent airflow, central if respiratory effort was absent, and mixed if there was absent respiratory effort initially followed by resumption of respiratory effort [13]. A hypopnea was defined as a reduction in airflow signal by ≥ 30% of baseline lasting at least 10 s and associated with at least a 4% oxyhemoglobin desaturation from pre-event baseline [7, 13, 14]. The following variables were calculated for each sleep study: AHI; number of apneas; number of obstructive, central, and mixed apneas; and number of hypopneas.

The HSATs were scored separately using one of three different airflow signals: (a) nasal pressure (NP), (b) transformed (square root) nasal pressure signal (transformed NP), and (c) uncalibrated respiratory inductive plethysmography (RIP) flow [6, 15, 16]. Only one airflow signal was used for each scoring session. For example, when the NP was used as the signal for airflow, both the transformed NP and RIP flow were not visible to the scorer.

Automated scoring

Two commercially available automated software scoring systems were used: (i) Remlogic (RLG) and (ii) Noxturnal (NOX). The studies in EDF were imported into each commercially available software and two investigators (KAW, JNJ) blinded to the manual scoring results, ran the automated scoring systems. The same "lights out" and "lights on" times were used as in the manual scoring above. Each study was scored separately using one of the three different airflow signals as above and using the same definition for apneas and hypopneas as the manual scoring. Automated scoring was performed without additional manual human review of the respiratory events.

Sample size

The primary outcome for the inter-rater agreement analysis was the intra-class correlation coefficient (ICC) of the AHI. Given 11 scorers (9 humans and 2 automated scorers), the 15 PSGs had a power of 94% to detect an ICC of at least 0.90, assuming a null hypothesis of ICC = 0.70 and a type 1 error rate of 0.05.

Statistical analysis

The inter-rater reliability agreement among the 11 different scorers (9 humans and 2 automated scorers) was assessed using the ICCs for the different respiratory indices.

Additionally, we also determined the ICCs of the average of the manual scoring and the two automated systems. The levels of agreement using the ICCs of respiratory indices were classified as follows: 0.00–0.25 = little, 0.26–0.49 = low, 0.50–0.69 = moderate, 0.70–0.89 = strong, and 0.90–1.00 = very strong [17, 18]. We also calculated the mean difference ($MEAN_{diff}$) between the average of the manual scores and the automated scores of the AHI. The limits of agreement (LoA) were calculated as mean difference ± 1.96*SD of the differences [19]. Data analyses were performed using SPSS software version 23 (IBM Corp., Armonk, NY). The Bland-Altman plots were generated using SigmaPlot software version 13 (Systat Software, Inc., San Jose, CA).

Results

The AHI values of all 15 HSATs by the nine scorers and two automated systems using the three different airflow signals are shown in Fig. 1. The mean ± SD of the AHI on all the 15 HSATs was 27.8 ±5.6 events/h (range 6.3–105.5 events/h) using NP, 25.1 ± 2.2 events/h (range 4.2–73.4 events/h) using transformed NP, and 24.0 ±2.0 events/h (range 4.7–67.5 events/h) using RIP flow.

Scoring agreement assessed by intra-class correlation coefficients

The inter-scorer agreement of AHI scoring among the human and automated scorers was very strong using any of the following as the primary signal for airflow: NP, transformed NP, or RIP flow. As shown in Table 1, the ICCs of the respiratory event scoring using the NP were AHI = 0.96 [95% CI 0.93–0.99], apnea index = 0.87 [0.77–0.95], and hypopnea index = 0.67 [0.50–0.84]. The ICCs using the transformed NP were AHI = 0.98 [0.96–0.99], apnea index = 0.89 [0.81–0.96], and hypopnea index = 0.81 [0.68–0.92]. The ICCs using the RIP flow were AHI = 0.97 [0.95–0.99], apnea index = 0.66 [0.49–0.84], and hypopnea index = 0.79 [0.65–0.91]. Scoring agreements of the apnea and hypopnea indices were therefore not as high as the scoring agreement of the AHI, although agreement was still in the moderate to strong range. In addition, as shown in Table 1, there was degradation of the scoring agreement of central and mixed apneas.

When the average of the manual scoring was used to determine the inter-rater agreement with the two automated systems, similar results were obtained. As shown in Supplemental Table 1, the ICCs using the NP were AHI = 0.99 [0.97–0.99], apnea index = 0.78 [0.56–0.91], and hypopnea index = 0.52 [0.20–0.78]. The ICCs using the transformed NP were AHI = 0.98 [0.96–0.99], apnea index = 0.79 [0.57–0.91], and hypopnea index = 0.72 [0.47–0.88]. The ICCs using the RIP flow were AHI = 0.96 [0.91–0.99], apnea index = 0.71 [0.46–0.88], and hypopnea index = 0.90 [0.79–0.96].

Scoring agreement assessed by the mean difference

Scoring agreement was also determined using the mean difference ($MEAN_{diff}$) between the average of the manual scores and the automated scores of the AHI using the three airflow signals. As shown in Fig. 2, using the NP signal, the mean differences in AHI were – 0.9 ± 3.1/h (AHI_{RLG} VS AHI_{MANUAL}), –1.3±2.6/h (AHI_{NOX} VS AHI_{MANUAL}), and –0.3±1.9/h (AHI_{RLG} VS AHI_{NOX}). Overall, the mean differences in AHI scoring between manual and

automated scoring were small. In addition, the mean difference in AHI scoring between the two automated scoring systems was small. Table 2 shows the mean differences between manual and automated systems using the three different airflow signals.

Discussion

The current research investigated HSAT scoring agreement of respiratory events between manual scoring by international sleep technologists and two commercially available automated scoring systems. We found the following results: (1) there is a very strong inter-rater agreement in the scoring of the AHI for HSATs between the automated systems and manual scoring performed by technologists from international sleep centers; (2) the scoring agreement of the AHI between the automated scoring systems and experienced technologists is very strong regardless of the type of airflow signal used (NP, transformed NP, or RIP flow); (3) there is very strong agreement in the scoring of the AHI between the two commercially available automated scoring systems used in this study; and (4) scoring agreement of the apnea and hypopnea indices was not as high as the scoring agreement of the AHI, although agreement was still in the moderate to strong range. We are unaware of any previous study that examined the agreement of HSAT scoring of respiratory events among international sleep technologists and automated scoring systems. Two prior studies reported that automated scoring underestimates the apnea-hypopnea index (AHI) obtained by human scorers with a mean difference ranging from 5 to 8 events/h [8, 9]. However, both these studies compared the automated systems to only one human scorer. In contrast, the automated systems in our study were compared to nine human scorers which provided a more accurate comparison considering that manual scoring among sleep scorers has inherent variability. As shown in Fig. 1, the AHI derived from the two automated systems was within the variation of the AHI scored by the nine human sleep technologists in each of the 15 HSATs. In addition, the fact that different scoring software were utilized by the human scorers is a strength of our current study given that this would be the real-world situation in collaborative endeavors involving international sleep centers.

Three previous studies have examined the scoring agreement of in-laboratory PSGs between human and automated systems [20–22]. All of these studies were done in the USA and all reported good agreement in the scoring of respiratory events between experienced sleep technologists and the automated systems. In the largest of these studies, by Malhotra and colleagues, automated scoring compared to ten sleep technologists from five sleep centers also yielded very good ICCs for sleep architecture measures aside from the AHI. The authors suggested that automated scoring, particularly if supplemented with manual editing, may provide a way to improve sleep laboratory efficiency as well as standardize PSG scoring across sleep centers [21]. The current study extends the findings of these previous reports to scoring of HSATs and among international sleep centers. This has implications particularly in research, given the increasing use of HSAT to diagnose OSA and the developing collaboration among sleep researchers worldwide in multi-center international studies [10, 11].

The agreement in the scoring of central apneas for HSATs between the average of the manual scoring and the two automated systems was lower than the scoring agreement for

obstructive apneas. Our previous studies also showed that, among human scorers, there was greater disagreement in the scoring of central apneas and mixed apneas for PSGs and HSATs than for scoring of obstructive apneas [7, 14]. The explanation for these findings is not known, but may be related to difficulties in determining the absence of respiratory effort from the abdominal and thoracic belt signals. The airflow derived by the square root transformation of the nasal pressure (transformed NP) signal is used by some centers for the scoring of respiratory events [6, 23, 24]. The RIP flow from the thorax and abdominal belt signals is also suggested as an alternative sensor for scoring respiratory events if the NP signal fails. Our study showed strong agreement in AHI scoring between human and automated systems when using the transformed NP or RIP flow signals. To our knowledge, ours is the first study that has examined the AHI scoring agreement using these alternative flow signals between human scorers and automated systems.

Our study has some limitations. First, similar to the two previous studies examining the agreement between manual scoring and automated systems in HSATs, we used a ≥ 4% oxygen desaturation to define a hypopnea, which is an acceptable alternative definition, but not the recommended one according to the update of the American Academy of Sleep Medicine (AASM) PSG scoring rules which uses a ≥ 3% oxygen desaturation [6]. A second limitation is the fact that all 15 HSATs were acquired from one SAGIC center. However, the type 3 portable device used to acquire the signals in these 15 HSATs included standard signal channels that would not be expected to be different in other international centers. A third limitation of our study is that we only included two automated scoring systems and our results cannot be extrapolated to other commercially available automated systems. Finally, although the 15 HSATs used in our study were randomly selected, all had good signal quality which is not to be expected with all HSATs. In the real world, human scorers will still be needed to review the HSATs to assess the quality of the signal to check for signal loss or high levels of noise.

In summary, our findings show that there is a strong agreement in the scoring of the AHI for HSATs between the commercially available automated scoring systems and experienced technologists from international sleep centers. This is also true when alternative airflow signals such as the transformed NP and RIP flow are used by the automated systems in the scoring of HSATs. Scoring agreements of the apnea and hypopnea indices were not as high as the scoring agreements of the AHI, although agreement was still in the moderate to strong range. Our study suggests that automated scoring of HSATs may be a useful tool to improve sleep laboratory efficiency and to standardize scoring among international sleep centers.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the following individuals who helped in this project: Mohammad Ahmadi, Alexander Blau, Petra Cornell, Silverio Garbuio, Su-Lan Liu, Joao Reinfelderon, Beth Staley, Magdalena Ósk Sigurgunnarsdóttir, and Sandra Zimmermann.

Funding Supported by NHLBI award P01 HL094307 (AIP), Conselho Nacional de Desenvolvimento Cientffico e Tecnologico (CNPq) grant 309336/2017–1 (LRB), Conselho Nacional de Desenvolvimento Cientffico e Tecno1ogico (CNPq), Grant 401569/2016–0 (LRB), and Award grant number UL1TR001070 from the National Center for Advancing Translational Sciences. The sponsor had no role in the design or conduct of the research.

Abbreviations

AHI	apnea-hypopnea Index
EDF	European data format
HSAT	home sleep apnea testing
ICC	intra-class correlation coefficient
$MEAN_{diff}$	mean difference
NOX	Noxturnal
NP	nasal pressure
PSG	polysomnography
RIP	respiratory inductive plethysmography
RLG	Remlogic
SAGIC	Sleep Apnea Global Interdisciplinary Consortium

References

1. Masa JF, Corral J, Pereira R, Duran-Cantolla J, Cabello M, Hernandez-Blasco L, Monasterio C, Alonso A, Chiner E, Rubio M, Garcia-Ledesma E, Cacelo L, Carpizo R, Sacristan L, Salord N, Carrera M, Sancho-Chust JN, Embid C, Vazquez-Polo FJ, Negrin MA, Montserrat JM (2011) Effectiveness of home respiratory polygraphy for the diagnosis of sleep apnoea and hypopnoea syndrome. Thorax 66(7):567–573 [PubMed: 21602541]
2. Whittle AT, Finch SP, Mortimore IL, MacKay TW, Douglas NJ (1997) Use of home sleep studies for diagnosis of the sleep apnoea/hypopnoea syndrome. Thorax 52(12):1068–1073 [PubMed: 9516901]
3. Kuna ST, Badr MS, Kimoff RJ, Kushida C, Lee-Chiong T, Levy P, McNicholas WT, Strollo PJ, on behalf of the ATS/AASM/ACCP/ERS Committee on Ambulatory Management of Adults with OSA (2011) An official ATS/AASM/ACCP/ERS workshop report: research priorities in ambulatory management of adults with obstructive sleep apnea. Proc Am Thorac Soc 8(1):1–16 [PubMed: 21364215]
4. Masa JF, Corral J, Pereira R, Duran-Cantolla J, Cabello M, Hernandez-Blasco L, Monasterio C, Alonso A, Chiner E, Zamorano J, Aizpuru F, Montserrat JM, and the Spanish Sleep Network (2011) Therapeutic decision-making for sleep apnea and hypopnea syndrome using home respiratory polygraphy: a large multicentric study. Am J Respir Crit Care Med 184(8):964–971 [PubMed: 21737584]
5. Flemons WW, Littner MR, Rowley JA, Gay P, Anderson WMD, Hudgel DW, McEvoy RD, Loube DI (2003) 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 124(4): 1543–1579 [PubMed: 14555592]
6. Berry RB, Budhiraja R, Gottlieb DJ, Gozal D, Iber C, Kapur VK, Marcus CL, Mehra R, Parthasarathy S, Quan SF, Redline S, Strohl KP, Davidson Ward SL, Tangredi MM, American Academy of Sleep Medicine (2012) Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events. Deliberations of the sleep apnea

definitions task force of the American Academy of Sleep Medicine. J Clin Sleep Med 8(5):597–619 [PubMed: 23066376]

7. Magalang UJ, Arnardottir ES, Chen NH, Cistulli PA, Gislason T, Lim D, Penzel T, Schwab R, Tufik S, Pack AI, SAGIC Investigators (2016) Agreement in the scoring of respiratory events among international sleep centers for home sleep testing. J Clin Sleep Med 12(1):71–77 [PubMed: 26350603]
8. Aurora RN, Swartz R, Punjabi NM (2015) Misclassification of OSA severity with automated scoring of home sleep recordings. Chest 147(3):719–727 [PubMed: 25411804]
9. Masa JF, Corral J, Pereira R, Duran-Cantolla J, Cabello M, Hernandez-Blasco L, Monasterio C, Alonso-Fernandez A, Chiner E, Vazquez-Polo FJ, Montserrat JM, the Spanish Sleep Group (2013) Effectiveness of sequential automatic-manual home respiratory polygraphy scoring. Eur Respir J 41(4):879–887 [PubMed: 22878873]
10. Hedner J, Grote L, Bonsignore M, McNicholas W, Lavie P, Parati G, Sliwinski P, Barbe F, de Backer W, Escourrou P, Fietze I, Kvamme JA, Lombardi C, Marrone O, Masa JF, Montserrat JM, Penzel T, Pretl M, Riha R, Rodenstein D, Saaresranta T, Schulz R, Tkacova R, Varoneckas G, Vitols A, Vrints H, Zielinski J (2011) The European Sleep Apnoea Database (ESADA): report from 22 European sleep laboratories. Eur Respir J 38(3):635–642 [PubMed: 21622583]
11. McEvoy RD, Antic NA, Heeley E, Luo Y, Ou Q, Zhang X, Mediano O, Chen R, Drager LF, Liu Z, Chen G, du B, McArdle N, Mukherjee S, Tripathi M, Billot L, Li Q, Lorenzi-Filho G, Barbe F, Redline S, Wang J, Arima H, Neal B, White DP, Grunstein RR, Zhong N, Anderson CS, SAVE Investigators and Coordinators (2016) CPAP for prevention of cardiovascular events in obstructive sleep apnea. N Engl J Med 375(10):919–931 [PubMed: 27571048]
12. Kemp B, Varri A, Rosa AC, Nielsen KD, Gade J (1992) A simple format for exchange of digitized polygraphic recordings. Electroencephalogr Clin Neurophysiol 82(5):391–393 [PubMed: 1374708]
13. Berry RB, Brooks R, Gamaldo C, et al. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications. Version 2.4 In: Darien. Version 2.4. In: Darien, IL: American Academy of Sleep Medicine; 2017
14. Magalang UJ, Chen N-H, Cistulli PA et al. (2013) Agreement in the scoring of respiratory events and sleep among international sleep centers. Sleep 36(4):591–596 [PubMed: 23565005]
15. Dingli K, Coleman EL, Vennelle M, Finch SP, Wraith PK, Mackay TW, Douglas NJ (2003) Evaluation of a portable device for diagnosing the sleep apnoea/hypopnoea syndrome. Eur Respir J 21(2): 253–259 [PubMed: 12608438]
16. Smith LA, Chong DW, Vennelle M, Denvir MA, Newby DE, Douglas NJ (2007) Diagnosis of sleep-disordered breathing in patients with chronic heart failure: evaluation of a portable limited sleep study system. J Sleep Res 16(4):428–435 [PubMed: 18036089]
17. Munro B Statistical methods for health care research. 5th ed Philadelphia: Lippincott Williams Wilkins; 2005
18. Cheng JW, Tsai WC, Yu TY, Huang KY Reproducibility of sonographic measurement of thickness and echogenicity of the plantar fascia. J Clin Ultrasound 40(1):14–19 [PubMed: 22109854]
19. Bland JM, Altman DG (1999) Measuring agreement in method comparison studies. Stat Methods Med Res 8(2):135–160 [PubMed: 10501650]
20. Pittman SD, MacDonald MM, Fogel RB et al. (2004) Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing. Sleep 27(7):1394–1403 [PubMed: 15586793]
21. Malhotra A, Younes M, Kuna ST, Benca R, Kushida CA, Walsh J, Hanlon A, Staley B, Pack AI, Pien GW (2013) Performance of an automated polysomnography scoring system versus computer-assisted manual scoring. Sleep 36(4):573–582 [PubMed: 23565003]
22. Punjabi NM, Shifa N, Dorffner G, Patil S, Pien G, Aurora RN (2015) Computer-assisted automated scoring of polysomnograms using the Somnolyzer system. Sleep 38(10):1555–1566 [PubMed: 25902809]
23. Thurnheer R, Xie X, Bloch KE (2001) Accuracy of nasal cannula pressure recordings for assessment of ventilation during sleep. Am J Respir Crit Care Med 164(10 Pt 1):1914–1919 [PubMed: 11734446]

24. Farre R, Rigau J, Montserrat JM, Ballester E, Navajas D (2001) Relevance of linearizing nasal prongs for assessing hypopneas and flow limitation during sleep. Am J Respir Crit Care Med 163(2):494–497 [PubMed: 11179129]

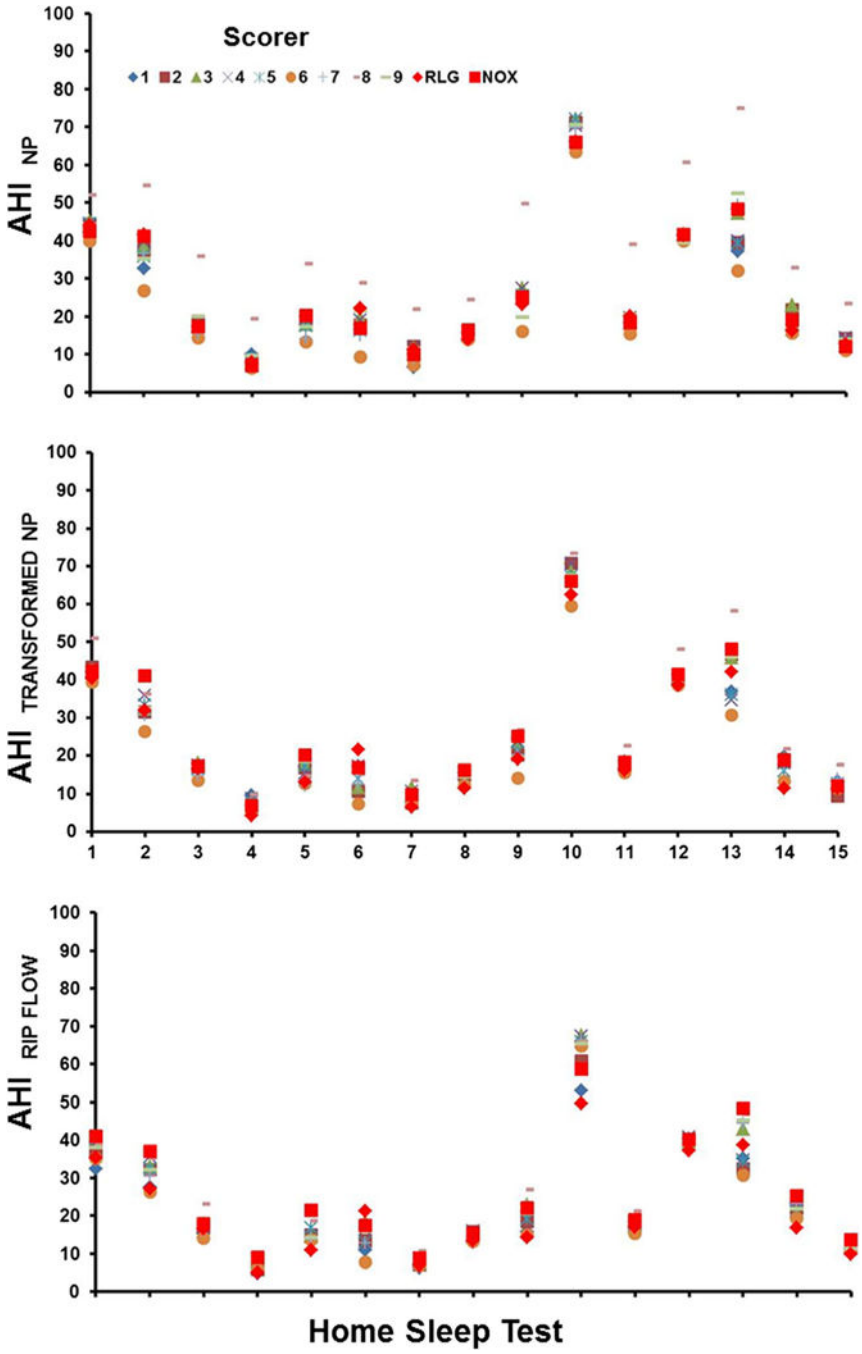

Fig. 1.

The absolute values of the AHI by the 9 human scorers and 2 automated systems of each of the 15 HSATs are shown. NP, nasal pressure; transformed NP, square root transformation of NP signal; RIP flow, flow using respiratory inductive plethysmography; AHI, apnea-hypopnea index; RLG, Remlogic; NOX, Noxturnal

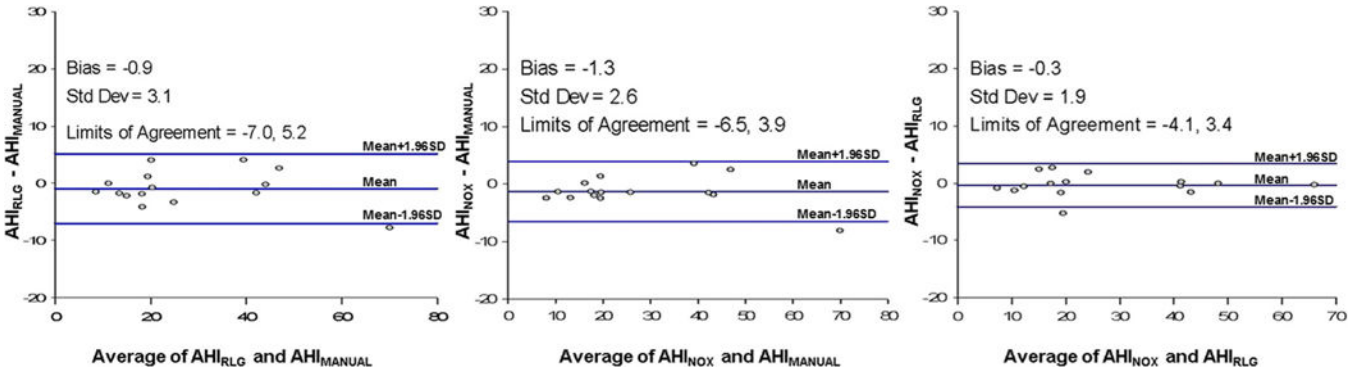

Fig. 2.

Bland-Altman plots of the agreement between the AHI values using the nasal pressure signal. AHI, apnea-hypopnea index; RLG, Remlogic; NOX, Noxturnal

Table 1

Inter-rater agreement of scoring respiratory events using different airflow signals. Values in parentheses represent the 95% confidence limits of the agreement among the 11 scorers (9 human and 2 automated scoring systems)

Variable	NP	Intra-class correlation coefficients Transformed NP	RIP flow
AHI (events/h)	0.96 [0.93–0.99]	0.98 [0.96–0.99]	0.97 [0.95–0.99]
Apnea index	0.87 [0.77–0.95]	0.89 [0.81–0.96]	0.66 [0.49–0.84]
OA index	0.86 [0.75–0.94]	0.73 [0.48–0.89]	0.60 [0.42–0.80]
CA index	0.59 [0.41–0.79]	0.80 [0.66–0.91]	0.61 [0.43–0.80]
MA index	0.56 [0.37–0.77]	0.66 [0.49–0.84]	0.42 [0.24–0.67]
Hypopnea index	0.67 [0.50–0.84]	0.81 [0.68–0.92]	0.79 [0.65–0.91]

NP, nasal pressure; *transformed NP*, square root transformation of NP signal; *RIP flow*, flow using respiratory inductive plethysmography; *AHI*, apnea-hypopnea index; *OA*, obstructive apnea; *CA*, central apnea; *MA*, mixed apnea

Table 2

Agreement between the average AHI values obtained from manual scoring by the 9 human scorers and the automated scoring systems

	NP			Transformed NP			RIP flow		
	$MEAN_{diff}$	**SD_{diff}**	**LoA**	**$MEAN_{diff}$**	**SD_{diff}**	**LoA**	**$MEAN_{diff}$**	**SD_{diff}**	**LoA**
AHI_{RLG} VS AHI_{MANUAL}	− 0.9	3.1	− 7.0, 5.2	−1.9	3.3	− 8.5, 4.6	−2.7	4.5	−11.6, 6.1
AHI_{NOX} VS AHI_{MANUAL}	− 1.3	2.6	− 6.5, 3.9	1.6	3.0	− 4.2, 7.4	2.3	3.4	− 4.4, 8.9
AHI_{NOX} VS AHI_{RLG}	− 0.3	1.9	− 4.1, 3.4	3.6	3.4	− 3.1, 10.2	5.3	4.2	− 3.2, 13.2

$MEAN_{diff}$, mean difference; *LoA*, limits of agreement; *NP,* nasal pressure; *transformed NP,* square root transformation of NP signal; *RIP flow,* flow using respiratory inductive plethysmography; *AHI,* apnea-hypopnea index