http://dx.doi.org/10.5665/sleep.2550

Agreement in Computer-Assisted Manual Scoring of Polysomnograms across Sleep Centers

Samuel T. Kuna, MD[1,2]; Ruth Benca, MD[3]; Clete A. Kushida, MD, PhD[4]; James Walsh, MD[5]; Magdy Younes, MB, ChB, PhD[6]; Bethany Staley, RPSGT[1]; Alexandra Hanlon, PhD[1]; Allan I. Pack, MB, ChB, PhD[1]; Grace W. Pien, MD[1]; Atul Malhotra, MD[7]

[1]Department of Medicine and Center for Sleep and Circadian Neurobiology, University of Pennsylvania, Philadelphia, PA; [2]Department of Medicine, Philadelphia VA Medical Center, Philadelphia, PA; [3]Department of Medicine, University of Wisconsin at Madison, Madison, WI; [4]Department of Psychiatry, Stanford University, Palo Alto, CA; [5]Sleep Medicine and Research Center, St. Luke's Hospital, Chesterfield, MO; [6]Department of Medicine, University of Manitoba, Winnipeg, Manitoba, Canada; [7]Department of Medicine, Harvard University, Boston, MA

Study Objectives: To determine intersite agreement in respiratory event scoring of polysomnograms (PSGs) using different hypopnea definitions.
Design: Technical assessment.
Setting: Five academic medical centers.
Participants: N/A.
Interventions: N/A.
Measurements and Results: Seventy good-quality PSGs performed in middle-aged women were manually scored by two experienced technologists at each of the five sleep centers using the particular laboratory's own software system. Studies were scored once by each scorer using American Academy of Sleep Medicine (AASM) standards for scoring sleep stages, arousals, and apneas. Hypopneas were then scored using three different AASM criteria: recommended, alternate, and research (Chicago). Means of each PSG variable for the scorers at each site were used to calculate an across-site intraclass correlation coefficient (ICC). Average AHI across the 10 scorers was 7.4 ± 12.3 (standard deviation) events/h using recommended criteria (ICC 0.984; 95% confidence interval [CI] 0.977-0.990), 12.1 ± 13.3 events/h using alternate criteria (ICC 0.947; 95% CI 0.889-0.972), and 15.1 ± 13.9 events/h with Chicago criteria (ICC 0.800; 95% CI 0.768-0.828). ICC across sites was 0.870 (95% CI = 0.847-0.889) for total sleep time, 0.861 (95% CI 0.837-0.881) for number of obstructive apneas and 0.683 (95% CI 0.640-0.722) for number of central apneas. ICCs across sites for hypopneas were very good using recommended criteria (ICC 0.843; 95% CI 0.820-0.870) but decreased when alternate criteria (ICC 0.728; 95% CI 0.689-0.763) and Chicago criteria (ICC 0.535; 95% CI 0.485-0.583) were used.
Conclusion: Experienced scorers at different laboratories have very good agreement in hypopnea and AHI results when good-quality PSGs are scored using AASM-recommended criteria. Substantial degradation of reliability was observed for alternative definitions of hypopneas, particularly that proposed for research.
Keywords: Apnea-hypopnea index, polysomnography, reliability, scoring
Citation: Kuna ST; Benca R; Kushida CA; Walsh J; Younes M; Staley B; Hanlon A; Pack AI; Pien GW; Malhotra A. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. *SLEEP* 2013;36(4):583-589.

INTRODUCTION

One of the known limitations of the polysomnogram (PSG) is its use of qualitative signals to determine sleep staging and identify abnormal respiratory events. Although biologic calibrations are obtained at the beginning and end of the PSG by having the patient perform various maneuvers, scoring of the recordings is largely based on pattern recognition. Although the PSG is a physiologic recording, its interpretation is more analogous to interpreting a radiographic image than to interpreting a pulmonary function test. Two other factors confound this problem. First, the sensors and computer systems used across laboratories to perform the recordings are not standardized. Even the performance of pulse oximeters differs from one manufacturer to another.[1,2] Second, several different criteria are used to score hypopneas,[3,4] and the apnea-hypopnea index (AHI), the PSG measure used to diagnose sleep apnea, can vary widely depending on the scoring criteria that is used.[5,6] As a result of these problems, sleep specialists in one sleep center may be uncertain how to interpret PSG reports generated by another sleep center. Multisite clinical sleep research studies can be designed to overcome this problem by having one sleep center score all of the studies performed across the sites.[7-9] Although centralized scoring decreases potential site-related differences, it is time consuming and requires funding of a separate core PSG laboratory with complicating issues related to transmission of files and sharing of private health information.

Previous investigators have reported that the level of agreement in sleep stage scoring of PSG is lower between laboratories than what can be maintained between scorers within the same laboratory.[10,11] Although Collop[12] found significant intersite variability in scoring of both sleep and respiratory events when the scoring criteria were not standardized across the sites, Magalang et al.[13] recently reported strong agreement in the scoring of the overall AHI as well as the number of apneas and hypopneas across the clinical centers when the sites all used the American Academy of Sleep Medicine (AASM)-recommended criteria for scoring respiratory events. The primary purpose of this study was to determine the level of agreement in respiratory event scoring of PSG across five sleep centers using each of three different definitions for hypopnea. We tested the hy-

A commentary on this article appears in this issue on page 465.

Submitted for publication June, 2012
Submitted in final revised form August, 2012
Accepted for publication August, 2012
Address correspondence to: Samuel T. Kuna, MD, Philadelphia VA Medical Center, 3900 Woodland Avenue, Philadelphia, PA 19104; Tel: (215) 823-4400; Fax: (215) 823-5194; E-mail: skuna@mail.med.upenn.edu.

Table 1—Criteria for scoring hypopneas

AASM recommended criteria[4]: Score a hypopnea if all of the following are present:

- Nasal pressure signal excursions (or those for the alternative hypopnea sensor) drop ≥ 30% of baseline for at least 10 sec
- There is a ≥ 4% desaturation from pre-event baseline
- At least 90% of the event's duration must meet the amplitude reduction of criteria for hypopnea

AASM alternative criteria[4]: Score a hypopnea if all of the following are present:

- Nasal pressure signal excursions (or those for the alternative hypopnea sensor) drop ≥ 50% of baseline for at least 10 sec
- There is a ≥ 3% desaturation from pre-event baseline or the event is associated with an arousal
- At least 90% of the event's duration must meet the amplitude reduction of criteria for hypopnea

AASM clinical research criteria (Chicago Criteria)[3]: Score a hypopnea if any of the following are present:

- Nasal pressure signal excursions (or those for the alternative hypopnea sensor) drop ≥ 50% of baseline for at least 10 sec
- The amplitude of the airflow or chest wall movement decreases to a level above 50% of the amplitude of "baseline" and is associated with a > 3% transient oxygen desaturation and/or an arousal.

AASM, American Academy of Sleep Medicine.

pothesis that the level of agreement in AHI across laboratories would depend primarily on the criteria used to score hypopneas and that the highest level of agreement would be present when scoring of hypopneas required an associated transient oxygen desaturation. The data were also used to determine if the level of agreement in AHI is lower across laboratories than what can be attained between scorers within the same laboratory. We anticipated that the level of agreement for scoring apneas would be higher than that for scoring hypopneas regardless of the criteria used to score hypopneas.

METHODS

Seventy PSGs were manually scored with the aid of computer software by two PSG technologists at the AASM-certified sleep centers at the University of Pennsylvania, the University of Wisconsin at Madison, St. Luke's Hospital in Chesterfield, MO, Stanford University, and Harvard University. All 10 scorers had at least 4 years of experience scoring PSGs and nine were registered PSG technologists. The recordings were selected from a set of deidentified home unattended overnight PSGs (Compumedics Safiro; Compumedics Sleep, Abbotsville, Australia) obtained during the baseline assessment (2007-2009) of an ongoing research study at the University of Pennsylvania examining sleep disordered breathing in women in midlife. The following inclusion criteria were used to enroll participants into that study: premenopausal, perimenopausal, and postmenopausal women between 40-57 y old with at least one ovary. Individuals were excluded from participation if they had any of the following: currently pregnant or breastfeeding, serious health problems (e.g., cancer), tracheostomy, current use of hormonal contraception or hormone replacement therapy, and hysterectomy. The parent study was approved by the Institutional Review Board (IRB) at the University of Pennsylvania and informed consent was obtained from all participants. As a part of that informed consent, participants approved the use of their deidentified data to be used in other research studies. The IRBs at three of the participating sites did not require a separate IRB approval to use the deidentified PSG files in this substudy. Additional IRB approval was required and obtained at the University of Wisconsin at Madison and Brigham and Women's Hospital in Boston, MA.

The techniques used to perform the PSGs were similar to those detailed previously.[8,14,15] The same montage, channel names, and sampling rates were used for all recordings. The following signals were recorded: electroencephalogram (C3M2 and C4M1), bilateral electrooculogram, chin muscle activity, and rib cage and abdominal excursion (piezoelectric crystal). Airflow was assessed by nasal airway pressure and oronasal thermistry. Body position, electrocardiogram (lead 1), and oxygen saturation (by pulse oximetry) were also recorded. The 70 PSGs were selected from a set of 222 PSGs from the baseline study assessment. Criteria for selection of the 70 PSGs included at least 5 h of scoreable data on the nasal pressure, thermistor, and effort channels, and a scoreable oxygen saturation signal for at least 95% of the study. The amount of sleep disordered breathing was not used in the selection process.

The PSGs with their unfiltered signals were converted to European Data Format (EDF) files and posted on a password-protected website shared by the five sleep centers.[16] The technologists at the five sleep centers copied the files from the website and imported them into the computer software program used at each particular site for manual scoring: University of Pennsylvania (Sandman 8.0, Natus Medical Inc., San Carlos, CA), Stanford University (Alice 2.7.43, Philips-Respironics, Murrysville, PA), Harvard University (Alice 2.7.43, Philips-Respironics), University of Wisconsin at Madison (Alice 2.7.43, Philips-Respironics), and St. Luke's Hospital (REMbrandt 7.5, Natus Medical Inc., San Carlos, CA). Because conversion of files into EDF does not retain events tagged manually on a recording, the scorers at each site were provided the start and end times of each PSG and the body position during the recording so these tags could be placed on the PSG files they scored.

The studies were scored once by each scorer using the current AASM standards for displaying the signals and scoring sleep stages, arousals, and apneas.[4,17] The studies were then scored three separate times by each scorer to mark the hypopneas and oxygen desaturation events, using three different scoring criteria for hypopneas: AASM-recommended, AASM alternate, and AASM clinical research (Chicago) criteria[3,4] (Table 1). The hypopneas and desaturations from the previous scoring passes were not visible during a subsequent scoring pass. The scorers were allowed to use their software's automatic scoring of oxygen desaturation events followed by manual review and editing. Mixed apneas were scored as obstructive. Each technologist scored all 70 studies using one of the three criteria for scoring hypopneas before scoring the studies with another scoring criteria. The order of the scoring method was randomized for each scorer. For each of the three scoring criteria, the studies were scored consecutively, according to the study identifica-

Table 2—Mean (SD) of polysomnogram measures across all scorers and intraclass correlation coefficients within and across sites

	Overall mean (SD) (10 scorers × 70 files)	Average SD of 70 files (range)	Average within-site ICC (range)	Across sites ICC (95% confidence interval)
Total sleep time (min)	397.6 (55.3)	16.2 (4.0-58.5)	0.892 (0.779-0.980)	0.870 (0.847-0.889)
Stage N1 (min)	43.3 (24.7)	16.3 (5.0-51.0)	0.618 (0.386-0.798)	0.441 (0.390-0.491)
Stage N2 (min)	241.9 (56.4)	31.6 (12.6-97.6)	0.752 (0.490-0.904)	0.614 (0.567-0.658)
Stage N3 (min)	33.0 (33.2)	22.6 (2.3-44.7)	0.557 (0.267-0.827)	0.402 (0.352-0.452)
Stage REM (min)	79.8 (29.2)	14.4 (4.1-41.1)	0.784 (0.636-0.915)	0.685 (0.642-0.724)
Sleep efficiency (%)	83.9 (8.6)	3.6 (0.9-10.6)	0.800 (0.653-0.961)	0.767 (0.731-0.798)
Latency to REM (min)	93.3 (48.7)	20.6 (0.2-131.2)	0.666 (0.319-0.897)	0.546 (0.496-0.593)
Arousals in REM (n)	19.9 (16.9)	10.4 (2.0-38.9)	0.547 (0.279-0.876)	0.520 (0.470-0.568)
Arousals in NREM (n)	89.5 (49.4)	29.3 (8.3-75.4)	0.594 (0.240-0.751)	0.575 (0.526-0.621)
Central apneas (n)	2.2 (5.3)	1.7 (0-15.6)	0.631 (0.257-0.949)	0.683 (0.640-0.722)
Obstructive apneas (n)	15.3 (38.9)	8.8 (0-72.0)	0.843 (0.733-0.914)	0.861 (0.837-0.881)
Average SpO_2 (%)	95.5 (0.1)	0.2 (0-0.7)	0.998 (0.992-1.00)	0.994 (0.990-0.996)
O_2 desaturations ≥ 4% (n)	45.0 (9.1)	9.4 (0-47.6)	0.971 (0.922-0.996)	0.902 (0.855-0.935)
% time SpO_2 < 90%	3.7 (0)	0.1 (0-0.8)	0.997 (0.984-1.00)	1.00

ICC, intraclass correlation coefficient; NREM, nonrapid eye movement; REM, rapid eye movement; SD, standard deviation.

tion from 1 to 70. The scorers were cautioned not to discuss any aspect of their scoring with the other scorers at or outside their sleep center. The scorers received no other instructions or training on how to score the PSGs. An additional analysis was performed using an automatic scoring software program developed by YST Limited (Winnipeg, Manitoba, Canada), and those results are reported in a separate study.[18] Once scored, the PSG outcome measures were exported to an Excel spreadsheet and submitted to the biostatistician for independent analyses.

Statistical Analysis

The mean and standard deviation (SD) of the 10 scores for each of the 70 files were calculated for each PSG variable of interest. The average of the 10 scores was used as the consensus score for each file. The SD of the 10 scorers was also calculated for each file. Thus, 70 SDs were generated for each variable. Interscorer variability is reported as the average of the 70 SDs, representing the average interscorer variability for that variable. Within-site intraclass correlation coefficients (ICC) for every PSG variable were calculated to compare the results of the two scorers at each site. The average and range of the five within-site ICCs were reported. To compare scoring across sites, the means of each PSG variable for the two scorers at each site were used to calculate an across-sites ICC for that variable. ICC was interpreted as follows: 0-0.2 indicates poor agreement: 0.3-0.4 indicates fair agreement; 0.5-0.6 indicates moderate agreement; 0.7-0.8 indicates strong agreement; and > 0.8 indicates almost perfect agreement. These values are arbitrary cutoffs, but similar to those used by Landis and Koch[19] for the kappa statistic.

RESULTS

In the 70 women (47% African American) who had PSGs, mean age was 51.1 ± 4.2 (SD) y (range 40.1-57.6), mean body weight 85.2 ± 22.9 kg (range 46.8-150.9), mean height 162.7 ± 6.8 cm (range 149.9-180.3), and mean body mass index 32.9 ± 9.2 kg/m^2 (range 17.1-57.4).

Table 2 shows the results for all measures excluding hypopneas. The average of each variable across the 70 studies and 10 scorers is shown in the first column, along with the SD. These averages and SDs are consistent with those observed in typical sleep laboratory referrals.

The second column in Table 2 shows the results of variability among the 10 scorers (inter-scorer variability). The average SD is the average of 70 SDs, with each representing the scoring variance in an individual file. The range of SDs among the 70 files is indicated in parentheses. The magnitude of interscorer variability differed greatly among files and scoring variables (Table 2, second column). For total sleep time, for example, the SD ranged from 4.0 to 58.5 min among the 70 files (average = 16.2 min). For almost all variables there were examples of large variability between scorers (see Table 2 values in parentheses). The outlying values (both the highest and lowest for each of the different variables) did not belong to a single or even few scorers. Eight of the 10 scorers contributed to the outliers.

The third column in Table 2 shows the scoring variability between two scorers within individual sites (within-site variability). The average and range (parentheses) of the five ICCs are shown for each variable. Within-site agreement was excellent for oxygen saturation variables and total sleep time, fairly good for sleep efficiency, time in Stage N2, time in rapid eye movement (REM) sleep, and number of obstructive apneas, but poor for the remaining variables with ICCs values in the 0.2 to 0.4 range in some sites.

The across-sites ICC for total sleep time, number of obstructive apneas, average oxygen saturation, oxygen desaturations ≥ 4%, and percent of time oxygen saturation < 90% were all ≥ 0.861, i.e., "very good".[19] Lower ICCs were observed for scoring of specific sleep stages and arousals (Table 2, last column). Across-sites ICCs were generally lower than within-site ICCs, particularly for the scoring of different sleep stages.

The AHI of each participant based on scoring by AASM-recommended criteria was averaged across all 10 scorers. The mean AHI for all 70 studies was 7.4 ± 12.3 events/h (mean AHI

range 5.5-9.3). Similar to the results of Ruehland et al.,[5] the mean AHI increased (12.1 ± 13.3 events/h) when the studies were scored using AASM alternate criteria and increased further (15.1 ± 13.9 events/h) when the Chicago criteria were used. All of the participants in whom sleep apnea was diagnosed by a particular scoring method had obstructive sleep apnea (OSA). However, the three scoring criteria resulted in differences in the percentage of individuals in whom sleep apnea was diagnosed and its severity (Figure 1). Of the 61 participants with no OSA and mild OSA (AHI < 15) when the studies were scored using the AASM-recommended criteria, 6 (9.8%) had moderate to severe OSA (AHI ≥ 15) using the AASM alternate criteria and 14 (23.0%) had moderate to severe OSA using the Chicago criteria. The within-site ICC for AHI ranged from 0.910-0.996 when scoring with the recommended criteria, 0.929-0. 974 when scoring with the alternate criteria, and 0.779-0. 961 with the Chicago criteria. The across-site ICC was excellent for the recommended criteria AHI (0.984; 95% confidence interval [CI] 0.977-0.990) and the alternate criteria AHI (0.947; 95% CI 0.889-0.972) but fell to 0.800 (95% CI 0.768-0.828) when AHI was calculated using the Chicago criteria.

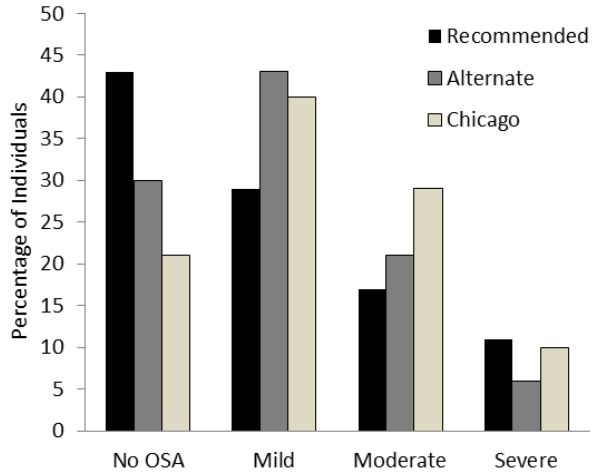

Figure 1—Percentage of patients without obstructive sleep apnea (no OSA; apnea-hypopnea index [AHI] < five events/h), mild OSA (five ≤ AHI < 15), moderate 15 ≤ AHI < 30), and severe OSA (AHI ≥ 30) using the American Academy of Sleep Medicine (AASM)-recommended, AASM alternate, and Chicago criteria. Results based on the average AHI using each of the three scoring criteria across all 10 scorers.

The variations in AHI across sites and across scoring criteria were primarily due to differences in the scoring of hypopneas (Table 3). Averaging across all 10 scorers, 29.1 ± 7.6 hypopneas were scored using the recommended criteria, 60.6 ± 19.5 with the alternate criteria, and 80.3 ± 38.4 with the Chicago criteria. The within-site ICC ranged from 0.699-0.953 scoring with the recommended criteria, 0.615-0.919 scoring with the alternate criteria, and 0.540-0.760 with the Chicago criteria (Table 4). The intersite ICC was 0.843 (95% CI 0.820-0.870) scoring with the recommended criteria, but progressively decreased with the alternate and Chicago criteria to 0.728 (95% CI 0.689-0.763) and 0.535 (95% CI 0.485-0.583), respectively (Table 4).

DISCUSSION

In agreement with the results of Ruehland et al.,[5] we found large differences in AHI depending on the AASM criteria used to score hypopneas. The mean AHI of the 70 PSGs scored by all 10 scorers across the five sleep centers was lowest when hypopneas were scored with the AASM-recommended criteria, higher when scoring used the AASM alternate criteria, and highest with the AASM Chicago criteria. The current results extend the findings of Ruehland et al.[5] by comparing intersite agreement using three different AASM-approved PSG scoring criteria. Very good intersite agreement (ICC ≥ 0.70) was found for total sleep time, number of OSAs, and all measures of oxygen saturation. ICCs across sites for hypopneas were very good when scored using AASM recommended criteria, but decreased progressively when alternate criteria and Chicago criteria were

Table 3—Mean (standard deviation) of total number of hypopneas scored by each scorer

Site	Scorer	Recommended	Alternate	Chicago
1	A	33.5 (44.4)	65.7 (62.6)	71.9 (63.2)
1	B	17.8 (23.3)	34.0 (32.3)	32.3 (33.3)
2	A	26.1 (37.1)	94.8 (63.1)	135.6 (82.5)
2	B	28.6 (38.0)	64.7 (53.2)	88.0 (64.3)
3	A	23.5 (28.9)	36.9 (32.9)	40.7 (36.4)
3	B	17.9 (25.6)	43.4 (41.5)	60.4 (51.5)
4	A	37.4 (44.2)	82.6 (63.0)	145.0 (86.3)
4	B	38.9 (54.6)	71.0 (66.7)	90.7 (75.7)
5	A	32.8 (48.4)	54.7 (53.8)	43.6 (51.4)
5	B	34.5 (45.9)	58.4 (53.7)	94.7 (57.1)
Mean of all scorers		30.2 (41.3)	59.1 (55.8)	86.9 (76.8)

Table 4—Intraclass correlation coefficients of total number of hypopneas within site and across site

	Recommended		Alternate		Chicago	
Criteria	ICC	95% of ICC	ICC	95% of ICC	ICC	95% of ICC
Site 1	0.699	0.396-0.839	0.615	0.134-0.815	0.569	-0.007-0.807
Site 2	0.953	0.925-0.971	0.776	0.177-0.915	0.660	0.088-0.854
Site 3	0.868	0.766-0.923	0.784	0.669-0.861	0.760	0.387-0.888
Site 4	0.942	0.908-0.963	0.919	0.838-0.956	0.718	-0.023-0.902
Site 5	0.922	0.877-0.950	0.881	0.815-0.924	0.540	-0.085-0.812
All sites	0.843	0.820-0.870	0.728	0.689-0.763	0.535	0.485-0.583

ICC, intraclass correlation coefficient.

used. As a result, the AHI across sites was remarkably similar when the PSGs were scored with the AASM recommended criteria (ICC 0.984) and only slightly less agreement in AHI (ICC 0.947) existed with the alternate criteria scoring.

The high intersite ICCs for oxygen desaturation events ≥ 4% (ICC = 0.902; range 0.855-0.935) helps explain why the agreement in scoring of hypopneas across sites was best when using AASM-recommended criteria. Those criteria require that a hypopnea be associated with at least a 4% oxygen desaturation event. The weaker agreement across sites for scoring arousals in nonrapid eye movement (NREM) and REM sleep (ICCs of 0.575 and 0.520, respectively) was associated with greater variability in scoring hypopneas using the AASM alternate and Chicago criteria, both of which allow scoring of hypopneas when the reduction in respiration is associated with an arousal. The variability found in scoring of arousals is in agreement with previous studies.[20-22] The lack of a metric calibration of the respiratory signals on PSG may account for the further decrease in intersite agreement using the Chicago criteria that permit a hypopnea to be scored with a ≥ 50% reduction in a respiratory signal even in the absence of an oxygen desaturation event or arousal. It is important to note, however, that piezoelectric crystals were used in this study to assess respiratory effort. It is possible that a better method of assessing effort, e.g., inductance plethysmography, might improve identification of hypopneas, particularly for the Chicago criteria. Furthermore, just because a particular scoring method results in the best agreement across scorers, i.e., highest ICC, does not necessarily mean that it is the best method in terms of clinical relevance. For example, given that many of the consequences of OSA result from sleep fragmentation, ignoring events that result in arousal but not desaturation may underestimate the clinical severity of the patient's disease and may, in fact, preclude the patient from getting needed treatment because of AHI criteria imposed by insurers and other payers. Therefore, the methods that give the best agreement may not be the best at assessing the severity of the patient's disease.

Previous studies have compared interscorer agreement of sleep stage and respiratory event scoring.[10-12,23,24] Whitney et al.[24] found a high degree of intrascorer and interscorer reliability for the scoring of sleep stage and respiratory disturbance index in 20 home unattended PSGs. The three scorers from the same laboratory were highly reliable for various respiratory disturbance indices, which incorporated an associated oxygen desaturation in the definition of respiratory events with or without the use of associated EEG arousal (ICC > 0.90). When respiratory disturbance index was defined without considering oxygen desaturation or arousal to define respiratory events, the respiratory disturbance index was moderately reliable (ICC = 0.74). Norman et al.[10] studied interobserver agreement in sleep stage scoring of 62 PSGs among five sleep scorers from different centers using the scoring criteria of Rechtschaffen and Kales.[25] The mean epoch-by-epoch agreement between scorers for all records was 73% (range 67%-82%). However, when considering the four stages of wake, nondelta, delta, and REM, the overall agreement obtained was 83% whereas the agreement for combined Stages N1 and N2 was 86%. Their epoch-to-epoch analysis showed an ICC of 0.51 for scoring Stage N1, 0.56 for scoring Stage N2, 0.71 for Stage N3-4, and 0.73 for Stage REM.Two other studies compared intersite scoring of sleep staging but did not study respiratory event scoring.[11,23] The intersite study of Danker-Hopfe et al.[23] used seven experienced PSG technologists at three sites to score sleep stages in 72 PSGs (56 healthy control patients and 16 patients with different sleep disorders) to compare the 2007 AASM criteria and the criteria of Rechtschaffen and Kales.[4,25] The relatively high interscorer ICCs for sleep staging may have been due to the 2-day training symposium held prior to scoring.

Collop[12] assessed scoring variability between PSG technologists in different sleep laboratories using their respective laboratory's scoring rules. Significant variability was present in scoring of both sleep and respiratory events with more variability demonstrated in respiratory event scoring. Most recently, Magalang et al.[13] evaluated the intersite agreement of scoring respiratory events and sleep stages in the nine center members of the Sleep Apnea Genetics International Consortium. One scorer at each site scored the same set of 15 PSGs using 2007 AASM-recommended criteria. The results were remarkably similar to those of the current study. Magalang et al.[13] found strong agreement in the scoring of the overall AHI as well as the number of apneas and hypopneas across the clinical centers. The intersite ICCs of the respiratory variables were AHI = 0.95 (95% CI 0.91-0.98), number of obstructive apneas = 0.69 (0.52-0.86), number of central apneas = 0.45 (0.26-0.69), number of hypopneas = 0.80 (0.66-0.91), and oxygen desaturation index = 0.97 (0.93-0.99). The intersite ICC of the arousal index was 0.68 (0.50-0.85). Similar to the agreement in sleep staging in our study using ICCs (Table 2), Magalang et al.[13] found substantial epoch-by-epoch intersite agreement in scoring wake, NREM, and REM sleep. The kappa statistics for sleep stages were: N1 = 0.31 (0.30-0.32), N2 = 0.60 (0.59-0.61), N3 = 0.67 (0.65-0.69), and REM = 0.78 (0.77-0.79). The current study extends the results of Magalang et al.[13] by showing that intersite agreement in scoring using AASM-recommended criteria is comparable to intrasite agreement, i.e., the agreement seen between scorers at the same site. In addition, we found that the intersite agreement progressively decreased when hypopneas were scored with the AASM alternative and Chicago criteria.

Our study and that of previous investigators have potentially important clinical implications.[13,23,24] When evaluating a patient who had a PSG at another sleep facility, a sleep provider can easily obtain the report of the PSG, but rarely has access to the actual recording. The inability to view the signals and scored events makes it difficult to rely on a sleep study report received from another facility. Uniform use of AASM-recommended criteria for scoring respiratory events by sleep centers and laboratories might help to partially alleviate this concern and allow the sleep specialist to interpret results from another laboratory with greater confidence. It is important to emphasize, however, that this study was done in five academic institutions using highly trained, experienced personnel who were aware that they were participating in a research project. Our results may therefore constitute a "best case" scenario and comparable results might not be obtained across clinical laboratories.

In agreement with Magalang et al.,[13] our results also indicate that centralized scoring of PSGs may not be necessary in multicenter research studies. Core sleep laboratories have been used in previous multicenter studies to reduce site-related differ-

ences in PSG scoring. Although some multisite studies such as the Sleep Heart Health Study[24] and Sleep AHEAD (Action for Health in Diabetes)[14] used the same equipment across participating sites to record the PSGs, the ability to acquire the recordings using each clinical center's existing equipment and software, whether portable or in-laboratory, would significantly reduce study-related cost. However, this introduces challenges related to differences in equipment across sites and transmission of the PSG files to the core laboratory for centralized scoring. Clinical centers participating in multicenter clinical trials are likely to be using different computer software programs to record and score PSGs. When a clinical center is using computer software that is different from that at the core scoring laboratory, the clinical center converts its PSG files to EDF files so they can be imported into the core laboratory's software for scoring.[16] This conversion process eliminates all tagged comments from the electronic file and can result in data corruption. Scorers at the core sleep laboratory must reenter the tags once the EDF file is imported into their software and need to remain wary of signal corruption during the course of the research project, especially at study inception and after implementation of PSG software upgrades at the clinical center and core laboratory.

The ability of each participating clinic center to score its own PSGs has the additional advantage of reducing the turnaround time for analysis. When the PSGs obtained for research protocols are part of a participant's usual clinical care, the analysis must be performed promptly so the results can be used to determine subsequent clinical management. Even with the use of Web portals to transmit the PSG files between clinical centers and the core sleep laboratory, the time required to transfer and process the files can add several days of delay in obtaining the results. The ability of each clinical site to score its own recordings would eliminate that delay. The use of the automated scoring software reported in our companion paper by Malhotra et al.[18] could further expedite the analysis process.

The ability of multisite research studies to have PSGs and home sleep tests scored by the individual sites using AASM criteria requires an ongoing quality control program at each clinical site and does not eliminate the need for establishing policies and procedures for performing the recordings and to standardize data collection. A quality assurance program should also be conducted to monitor scoring quality and ensure the sites' adherence to PSG-related policies and procedures. A core PSG laboratory's oversight of activities at the sites would be a significant cost savings compared to its scoring all of the studies.

In summary, we found very good agreement in the scoring of hypopneas across five academic sleep centers when scored by experienced PSG technologists according to AASM- recommended criteria. The intersite agreement in PSG outcome measures that are frequently used in multicenter clinical sleep research studies suggests that the scoring could be performed at the individual sites rather than centralized scoring by a core sleep laboratory. Adequate centralized supervision would still be required to standardize the recordings and monitor adherence to policies and procedures. However, allowing sites to score their own studies using AASM-recommended criteria would reduce the cost of the research and improve timeliness of results. Greater efforts are also needed to standardize PSGs in clinical care. Uniform adoption of the AASM-recommended criteria would also help to achieve the long standing but elusive goal of standardizing PSGs in clinical care.

ACKNOWLEDGMENTS

Scorers at the University of Pennsylvania: Haideliza Soto-Calderon, Mary Jones-Parker, RPSGT; scorers at Harvard University: Mary MacDonald, RPSGT, Pam DeYoung, RPSGT; scorers at the University of Wisconsin at Madison: Jennifer Noe, RPSGT, Rebecca Weise, RPSGT; scorers at Stanford University: Martin Ukockis, RPSGT, Julianne Blythe; scorers at St. Luke's Hospital: Hilary Field, RPSGT, Laura Rahubka, RPSGT. Eileen O'Leary, RPSGT coordinated activity at Stanford University and assisted in data export at the sites using Alice software. Sources of support: NIH R01 HL085695; NIH 1PO1-1HL094307; Academic Alliance for Sleep Research (Philips-Respironics Foundation).

DISCLOSURE STATEMENT

This was not an industry supported study. **Dr. Kuna** receives grant support from Philips-Respironics. **Dr. Benca** has received consulting income from Merck and Sanofi-Aventis. **Dr. Malhotra** had previously received (prior to May 2012) consulting and/or research income from Philips Respironics, SHC, SGS, Apnex Medical, Pfizer and Apnicure. **Dr. Kushida** receives research support from ResMed, Pacific Medico, Merck, Cephalon, Ventus Medical, and Jazz Pharmaceuticals. He receives royalties from Philips-Respironics. **Dr. Younes** is the owner of YRT a research and development company. **Dr. Walsh** has received research support from Apnex, NovoNordisk, Merck, Phillips-Respironics, Vanda, and Ventus. The other authors have indicated no financial conflicts of interest.

REFERENCES

1. Trivedi NS, Ghouri AF, Lai E, Shah NK, Barker SJ. Pulse oximeter performance during desaturation and resaturation: a comparison of seven models. J Clin Anesth 1997;9:184-8.
2. Zafar S, Ayappa I, Norman RG, Krieger AC, Walsleben JA, Rapoport DM. Choice of oximeter affects apnea-hypopnea index. Chest 2005;127:80-8.
3. American Academy of Sleep Medicine Task Force. Sleep-related breathing disorders in adults: Recommendations for syndrome definitions and measurement techniques in clinical research. Sleep 1999;22:667-89.
4. Iber C, Ancoli-Israel S, Chesson AL, Quan SF, for the American Academy of Sleep Medicine. The AASM Manual for the scoring of sleep and associated events. Westchester, IL: American Academy of Sleep Medicine; 2007.
5. Ruehland WR, Rochford PD, O'Donoghue FJ, Pierce RJ, Singh P, Thornton AT. The new AASM criteria for scoring hypopneas: impact on the apnea hypopnea index. Sleep 2009;32:150-7.
6. Redline S, Kapur V, Sanders MH, et al. Effects of varying approaches for identifying respiratory disturbances on sleep apnea assessment. Am J Respir Crit Care Med 2000;161:369-74.
7. Foster GD, Kuna ST, Sanders M, et al. Sleep apnea in obese adults with type 2 diabetes: baseline results from the Sleep AHEAD study. Sleep 2005;28:A205.
8. Redline S, Sanders MH, Lind BK, et al. Methods for obtaining and analyzing unattended polysomnography data for a multicenter study, Sleep Heart Health Research Group. Sleep 1998;21:759-67.
9. Kuna ST, Gurubhagavatula I, Maislin G, et al. Noninferiority of functional outcome in ambulatory management of obstructive sleep apnea. Am J Respir Crit Care Med 2011;183:1238-44.
10. Norman RG, Pal I, Stewart C, Walsleben JA, Rapoport DM. Interobserver agreement among sleep scorers from different centers in a large dataset. Sleep 2000;23:901-8.

11. Ferri R, Ferri P, Colognola RM, Petrella MA, Musumeci SA, Bergonzi P. Comparison between the results of an automatic and a visual scoring of sleep EEG recordings. Sleep 1989;12:354-62.
12. Collop NA. Scoring variability between polysomnography technologists in different sleep laboratories. Sleep Med 2002;3:43-7.
13. Magalang UJ, Chen N, Cistulli PA, et al. Agreement in the scoring of respiratory events and sleep amoung international sleep centers. Sleep 2013;36:591-6.
14. Foster GD, Sanders MH, Millman R, et al. Obstructive sleep apnea among obese patients with type 2 diabetes. Diabetes Care 2009;32:1017-9.
15. Rodway GW, Weaver TE, Mancini C, et al. Evaluation of sham-CPAP as a placebo in CPAP intervention studies. Sleep 2010;33:260-6.
16. Kemp B, Varri A, Rosa AC, Nielsen KD, Gade J. A simple format for exchange of digitized polygraphic recordings. Electroenceph Clin Neurophysiol 1992;82:391-3.
17. Silber MH, Ancoli-Israel S, Bonnet MH, et al. The visual scoring of sleep in adults. J Clin Sleep Med 2007;3:121-31.
18. Malhotra A, Younes M, Kuna ST, et al. Performance of an automated polysomnography scoring system vs. computer-assisted manual scoring. Sleep 2013;36:573-82.
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.
20. Drinnan MJ, Murray A, Griffiths CJ, Gibson GJ. Interobserver variability in recognizing arousal in respiratory sleep disorders. Am J Respir Crit Care Med 1998;158:358-62.
21. Loredo JS, Clausen JL, Ancoli-Israel S, Dimsdale JE. Night-to-night arousal variability and interscorer reliability of arousal measurements. Sleep 1999;22:916-20.
22. Ruehland WR, O'Donoghue FJ, Pierce RJ, et al. The 2007 AASM recommendations for EEG electrode placement in polysomnography: impact on sleep and cortical arousal scoring. Sleep 2011;34:73-81.
23. 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;18:74-84.
24. Whitney CW, Gottlieb DJ, Redline S, et al. Reliability of scoring respiratory disturbance indices and sleep staging. Sleep 1998;21:749-57.
25. Rechtschaffen A, Kales A. A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects. Los Angeles: Brain Information Service/Brain Research Institute, University of California; 1968.