JCSM
Journal of Clinical Sleep Medicine

SCIENTIFIC INVESTIGATIONS

Minimizing Interrater Variability in Staging Sleep by Use of Computer-Derived Features

Magdy Younes, MD[1,2,3]; Patrick J. Hanly, MD[2]

[1]YRT Ltd, Winnipeg, MB, Canada; [2]Sleep Centre, Foothills Medical Centre, Calgary, Alberta, Canada; [3]Sleep Disorders Centre, Winnipeg, Manitoba, Canada

Study Objectives: Inter-scorer variability in sleep staging of polysomnograms (PSGs) results primarily from difficulty in determining whether: (1) an electroencephalogram pattern of wakefulness spans > 15 sec in transitional epochs, (2) spindles or K complexes are present, and (3) duration of delta waves exceeds 6 sec in a 30-sec epoch. We hypothesized that providing digitally derived information about these variables to PSG scorers may reduce inter-scorer variability.

Methods: Fifty-six PSGs were scored (five-stage) by two experienced technologists, (first manual, M1). Months later, the technologists edited their own scoring (second manual, M2). PSGs were then scored with an automatic system and the same two technologists and an additional experienced technologist edited them, epoch-by-epoch (Edited-Auto). This resulted in seven manual scores for each PSG. The two M2 scores were then independently modified using digitally obtained values for sleep depth and delta duration and digitally identified spindles and K complexes.

Results: Percent agreement between scorers in M2 was 78.9 ± 9.0% before modification and 96.5 ± 2.6% after. Errors of this approach were defined as a change in a manual score to a stage that was not assigned by any scorer during the seven manual scoring sessions. Total errors averaged 7.1 ± 3.7% and 6.9 ± 3.8% of epochs for scorers 1 and 2, respectively, and there was excellent agreement between the modified score and the initial manual score of each technologist.

Conclusions: Providing digitally obtained information about sleep depth, delta duration, spindles and K complexes during manual scoring can greatly reduce interrater variability in sleep staging by eliminating the guesswork in scoring epochs with equivocal features.

Keywords: automated scoring, interobserver variability, PSG, sleep stages

Citation: Younes M, Hanly PJ. Minimizing interrater variability in staging sleep by use of computer-derived features. *J Clin Sleep Med* 2016;12(10):1347–1356.

INTRODUCTION

Interrater variability in scoring polysomnograms (PSGs) is an important problem in sleep medicine. For five-stage sleep scoring it is difficult to obtain agreement in > 85% of epochs, on average, even between expert scorers, and agreement can be as low as 50% in some PSGs.[1–12] In a recent analysis we found that inattention errors and scoring bias accounted for < 25% of the differences between highly experienced scorers.[12] The major reason was the presence of a large number of epochs that are difficult to classify such that a technologist may be willing to accept either of two, or even three, scoring options. Because scorers are obliged to make a decision, agreement (or lack thereof) in such epochs is left to chance. These "equivocal" epochs accounted for 28 ± 12% of all epochs on average and up to 76% of all epochs in individual PSGs.[12]

Most scoring difficulties involve one of three pairs of choices[1–12]: (1) awake (W) versus nonrapid eye movement (NREM) sleep, (2) NREM stages N1 versus N2, and (3) NREM stages N2 versus N3. W/NREM difficulties arise when the electroencephalogram (EEG) contains both awake and sleep patterns and it is difficult to decide which pattern occupies > 15 sec of a 30-sec epoch. Distinction between N1 and N2 is based on identification of spindles and K complexes.[13] Definition of K complexes is qualitative with no criteria for minimum amplitude or for the durations of the complex's different phases. In addition, manual scoring of spindles is subject to much inconsistency among scorers.[14,15] N2/N3 differences are clearly related to difficulty in correctly estimating delta wave duration in epochs with borderline delta wave prominence.

Digital methods are currently available to determine depth of sleep, to identify spindles and K complexes, and to calculate delta wave duration within epochs. The odds-ratio-product (ORP) is a continuous index of sleep/wake state with a range from zero (very deep sleep) to 2.5 (full wakefulness).[16] Transitional epochs receive an ORP score between 1.0 and 2.0.

BRIEF SUMMARY

Current Knowledge/Study Rationale: Inter-scorer variability in scoring polysomnograms results primarily from difficulty in determining whether: (1) an electroencephalogram pattern of wakefulness spans > 15 sec in transitional epochs, (2) spindles or K complexes are present, and (3) duration of delta waves exceeds 6 sec in a 30-sec epoch. We hypothesized that providing digitally derived information about these variables to the scorers may reduce inter-scorer variability.

Study Impact: Percent agreement between scorers improved dramatically after their scores were independently modified using digitally obtained information about sleep depth, delta duration, spindles, and K complexes. Provision of such information during scoring can greatly reduce interrater variability in sleep staging by eliminating the guesswork in scoring epochs with equivocal features.

Most W/NREM disagreements occur in this range.[16] Likewise, digital methods for identifying spindles and K complexes are available[17–25] and it is easy to digitally calculate delta wave duration in 30-sec epochs. We hypothesized that making this information available to technologists may help to unify the scoring of such equivocal epochs thereby reducing interrater scoring variability.

To test this hypothesis, initial disagreement in sleep staging between two experienced scorers was determined. Thereafter, scoring of each technologist was independently adjusted epoch-by-epoch such that if manual stage was NREM and ORP was > 1.5 (middle of the transitional zone) the sleep stage was changed to W, and *vice versa* (see Methods). Epochs staged as N1 were converted to N2 if a digitally identified spindle or K complex was present in the appropriate location, whereas a shift from manually scored N1 to N2 in the absence of digitally identified spindles or K complexes was overruled. Epochs staged N2 when digitally determined delta wave duration was > 6 sec were converted to N3, and *vice versa*. Agreements between the two modified scores, and between the modified score and the original manual score of each technologist, were subsequently determined. When the modified score of an epoch differed from the manual score of both technologists we determined whether it was assigned by the same scorers during five other scoring sessions of the same PSGs. Modified scores that were not seen in any other session were considered errors of the proposed approach.

METHODS

See "Expanded Methods" in supplemental material.

Fifty-six PSGs were randomly selected from the sleep centre's database at the University of Calgary to represent a broad spectrum of sleep pathology: severe obstructive sleep apnea (OSA) (apnea-hypopnea index (AHI) > 30, n = 8), moderate OSA (AHI 15–30, n = 10), mild OSA (AHI 5–15, n = 10), central sleep apnea (AHI > 15, n = 4), severe OSA on continuous positive airway pressure (CPAP) throughout (n = 5), periodic limb movement (PLM) disorder (PLM index > 25, n = 4), insomnia (n = 5), narcolepsy (n = 5), no sleep pathology (n = 5). These were the same PSGs used to validate the ORP[16] and to determine the reasons for interrater variability.[12] The PSGs included two central (C3/A2, C4/A1) and one occipital (O2/A1) EEG signals, two electro-oculograms, chin electromyogram (EMG), electrocardiogram and signals from chest and abdomen bands (Respitrace, Ambulatory Monitoring, Ardsley, NY, USA), nasal pressure and oronasal thermister, oxyhemoglobin saturation, and a microphone. They were recorded with a Sandman system (Natus Medical, Pleasanton, CA). The scoring performed for the initial clinical evaluation was not considered here.

The PSG files were manually scored by two certified PSG technologists, scorer 1 and scorer 2, each with > 10 y of experience, one from the Sleep Centre at the University of Calgary and one from the Sleep Centre at the University of Manitoba (manual 1). The two technologists did not work together previously. PSGs were mailed from Calgary to Winnipeg to be scored by scorer 2 who used the same type of Sandman viewer with the same filter and resolution settings. Several months later scorer 1 and scorer 2 were asked to review their own scoring and correct any errors (manual 2). The PSGs were then exported in the European Data Format (EDF) with no added filters and the EDF files were automatically scored (Auto) using a validated[9,26] automatic system (Michele Sleep Scoring [MSS], Winnipeg, Manitoba, Canada). Scorer 1, scorer 2 and a third senior scorer (scorer 3) were asked to edit the automatic score, epoch by epoch, and correct any score they disagreed with. Accordingly, there were seven manual scores for each epoch in each PSG, three each from scorer 1 and scorer 2, and one from scorer 3.

Following auto-scoring, Excel files were generated that listed average ORP and average delta duration in each 30-sec epoch and the location of each spindle and K complex identified by the program. A brief description of the method of calculating ORP is provided in the following paragraphs. For more details on the method of obtaining ORP and the other variables please see the supplemental material.

Fast Fourier transform is performed on the EEG in consecutive 3-sec epochs. The spectral pattern of the EEG in each epoch is assigned a four-digit number based on the relative powers in four frequency bands.[16] ORP is obtained from a look-up table that contains the probability of each of these 10,000 patterns occurring in epochs staged awake by a consensus of highly experienced technologists. The likelihood ranges from 100% (invariably seen during epochs staged awake) to zero (never seen during epochs staged awake). The final ORP value is the normalized probability (2.5 = 100% and 0 = 0%) of the assigned pattern to occur during wakefulness.[16]

Epoch-by-epoch sleep scoring of scorer 1 and scorer 2 in the second manual session (manual 2) was added to the excel sheets containing the 30-sec ORP values, total delta wave duration, and epochs with spindles and K complexes. Spindles and K complexes were distinguished as to whether they were found in the first or last 15 sec of the epoch. The scores of scorer 1 and scorer 2 were then modified individually, without regard to the scores of the other technologist. The modified scorer 1 and scorer 2 scores were placed in new columns. The following modifications were implemented in sequence starting from the beginning of the file:

1. If manual stage is NREM sleep (any stage) and average ORP is ≥ 1.5, stage was changed to W. If stage is W when average ORP is < 1.5, it was changed to N1.
2. If manual stage was N1 and there were one or more spindles or K complexes in the first half of the epoch or in the second half of the preceding epoch, the stage was converted to N2 and the change was carried forward until the manual stage was no longer N1 or four epochs elapsed without a spindle or K complex. When manual stage changed from N1 in one epoch to N2 in the next in the absence of a digitally identified, appropriately located spindle or K complex, the sleep stage was changed to N1 and the change was carried forward until a spindle or K complex was found or manual stage changed to any stage other than N2.

Spindles and K complexes occurring in a previous epoch were ignored if the previous epoch was staged awake in view of the results of specificity analysis that showed false positive identification of these events in stage awake (see Analyses section).

3. Finally, epochs staged N2 when delta wave duration was > 6 sec were converted to N3, and *vice versa*.

Epochs staged manually as rapid eye movement (REM) sleep were not modified.

Management of Manually Scored Arousals

Because NREM sleep stage is changed to N1 following arousals,[13] some N1/N2 discrepancies in the manual scoring were related to differences in the manual scoring of arousals. On average, scorer 1 scored 105.2 ± 58.3 arousals per PSG whereas scorer 2 scored 59.1 ± 36.2. Only 41.1 ± 27.0 arousals per PSG were common to both scorers. We elected not to alter the manually scored arousals because there are no agreed-upon guidelines for digital scoring of arousals. Accordingly, if one technologist scored an arousal followed by stage N1 while the other did not score an arousal and the epoch remained as N2, no intervention was made unless a digitally scored spindle or K complex occurred after the arousal and before the middle of the epoch, as the second modification in the previous paragraph. However, because of the inconsistency of arousal scoring, when an epoch was changed from N1 to N2, because there were spindles or K complexes, extension of N2 forward was not stopped when an arousal was manually scored. Such a practice, which we tried initially, created too many N1/N2 discrepancies that did not exist before. Accordingly, we carried forward the change from N1 to N2 through manually scored arousals unless 4 epochs elapsed without spindles or K complexes.

Analyses

Analysis to Determine Specificity of the Digitally Identified Spindles and K Complexes

To confirm that spindles and K complexes are specific to stage N2, the frequency of spindles and K complexes (number/30-sec epoch) was calculated in epochs staged unanimously (i.e., in all seven manual scores) as awake, N1, N2, N3, and REM sleep. Each spindle or K complex identified in epochs unanimously staged N1 sleep was visually inspected. In performing this analysis, we were surprised by the small number of epochs scored unanimously as stage N1 or N3 relative to the total number of epochs assigned these stages in at least one scoring session. Accordingly, we performed a systematic evaluation of the probability of a sleep stage scored by any technologist in one session being assigned in the six other scoring sessions.

Analyses to Determine the Effect of Proposed Approach on Scoring Results

The original and modified scores of scorer 1 and scorer 2 were copied to a new excel sheet. The scores of scorer 1 and scorer 2 in manual 1 and the three post-Auto scores (scorers 1–3) were added to the table (total of nine columns). The following calculations were made for each PSG:

1. Epoch-by-epoch % agreement between scorer 1 and scorer 2 in the original manual 2 scoring.
2. Epoch-by-epoch % agreement between modified scorer 1 and modified scorer 2. The change in % agreement reflects the overall benefit of the proposed protocol in reducing interrater variability.
3. Agreement (intraclass correlation) between scorer 1 and scorer 2 in total sleep time, and times in W, N1, N2, N3, and REM sleep before and after modification of their scoring.
4. Number of epochs where stage assigned by scorer 1 was changed to the stage assigned by scorer 2.
5. Number of epochs where stage assigned by scorer 2 was changed to the stage assigned by scorer 1. The sum of the last two values (4 and 5) represents a positive outcome in that the score of either highly qualified technologist should be acceptable.
6. Number of epochs where the manual stage of either or both scorers was changed to a completely different stage. These are potential errors resulting from the proposed protocol. To determine whether the third score was an acceptable score given the equivocal nature of the epoch, we scanned the other scores given by the three scorers during the five other manual exercises (two from manual 1 and three manual edits of Auto). Epochs where the new score was not seen in any of the other sessions were considered as true errors. These were further categorized as to the type of error (e.g., W wrongly called N1, N1 wrongly called N2, etc.).

All epoch numbers obtained in steps 4–6 were expressed as % of total epochs in the PSG. Unless otherwise indicated, results are given as mean ± standard deviation and 10–90% confidence interval in the 56 PSGs.

RESULTS

As reported previously,[12] patients were 35 females and 21 males, 51 ± 14 y in age. Body mass index was 35 ± 12 kg/m^2. Total sleep time was 244 ± 106 min. For the entire group AHI was 21 ± 25 h^{-1}, arousal/awakening index was 25 ± 14 h^{-1}, and PLM index was 17 ± 31 h^{-1}.

Stage Specificity of the Digitally Identified Spindles and K Complexes

Frequency of spindles and K complexes when sleep stage was the same in all seven sessions (unanimous agreement) is shown in **Figure S1** in the supplemental material. Spindle frequency was highest in epochs unanimously staged N2 sleep (2.91 ± 1.86/epoch), but was highly variable between patients (10–90%, 0.60–5.51/epoch), and was lowest in REM sleep (0.70 ± 0.76; 0.08–1.45/epoch). Spindle frequency in stage N3 was nearly half the frequency in stage N2. Many spindles were identified in stage W in the absence of visually identifiable spindles (1.33 ± 0.93; 0.34–2.58/epoch). Frequency of spindles in unanimous stage N1 was 1.14 ± 0.72/epoch (0.34–2.10/epoch).

A total of 892 spindles were digitally scored during epochs staged unanimously as N1 sleep. Visual inspection of these events revealed that only 43 (4.8%) had no visual correlate and could be considered as true errors that could influence results of the proposed approach. The remainder were consistent with a score of N1 in that: (1) 431 (48.3%) occurred in the second half of the epoch and were followed by a manual stage change to N2 sleep in the following epoch, (2) 291 (32.6%) were digitally identified during or before manually scored arousals in the first half of the epoch, (3) 76 (8.5%) were questionable by visual inspection because of borderline duration (≈ 0.5 sec) or because the event could easily be scored as a subthreshold arousal (arousal < 3 sec in duration), (4) 29 (3.3%) occurred during an extension of the awake pattern from a preceding epoch staged awake, and 22 (2.5%) were assigned digitally to the first half of the epoch because their onset was immediately before the 15-sec point but their body was mostly in the second half. Accordingly, they visually appeared to be in the second half.

K complexes were much less frequent than spindles (**Figure S1**; note the 10:1 scale difference). Their frequency was also maximal in epochs unanimously staged N2 sleep but there were relatively much fewer K complexes than spindles in stages W and REM sleep. Their frequency in epochs unanimously staged as N1 sleep was 0.11 ± 0.13/epoch. Of these (110 in total), 58.5% occurred in the second half of the epoch and were followed by a change in manual stage to N2, 35.6% were K arousals and in the rest (5.9%) an independent arousal started prior to the midpoint of the epoch thereby precluding a manual stage change to N2 sleep.[13]

Effect of the Proposed Approach on Interrater Variability

There were 40,260 epochs in the dataset. Before modifications, overall % agreement between the two scorers in manual 2 for five-stage scoring was 78.9% and kappa statistic was 71.1% (**Figure S2** in the supplemental material). Percent agreement for individual PSGs was 78.1 ± 9.7% (confidence interval [CI]: 64.8–89.8%). Most differences were between W and NREM sleep (46 ± 38; CI: 10–140 epochs/PSG), between N1 and N2 (42 ± 36; CI: 11–90 epochs/PSG) and between N2 and N3 (52 ± 47; CI: 0–142 epochs/PSG). After the modifications (**Figure S2**), % agreement for the entire dataset increased to 96.5% and kappa statistic increased to 95.1%. Percent agreement for individual PSGs was 96.5 ± 2.6% (CI: 92.3–99.8%). Nearly all remaining disagreements were between REM sleep and other stages and a few N1/N2 disagreements (10 ± 10; CI: 1–34 epochs/PSG). There was no correlation between the arousal/awakening index in the PSG and magnitude of difference between the two scorers in any sleep stage or in the magnitude of reduction in these disagreements following the correction.

Figures 1 and **2** illustrate the agreement between the two scorers in times spent in different stages before and after modifications. Except for agreement in REM time (**Figure 2**), which was not modified, all relations improved substantially with intraclass correlation coefficients (ICCs) increasing to ≈ 1.00 in every case.

Table 1 shows average values of sleep variables obtained by scorer 1 and scorer 2 before and after modifications. Scorer 2 scored more awake, N2, and REM times and less N1, N3, and total sleep times than scorer 1. These differences were highly significant. With the exception of REM time, which did not change as expected, differences between scorer 1 and scorer 2 disappeared after modifications. Modified N2 was higher, and N1 time was lower than the amounts scored by both technologists, mainly due to more epochs converted from N1 to N2 than the opposite. Modified N3 was intermediate between the values scored by the two technologists.

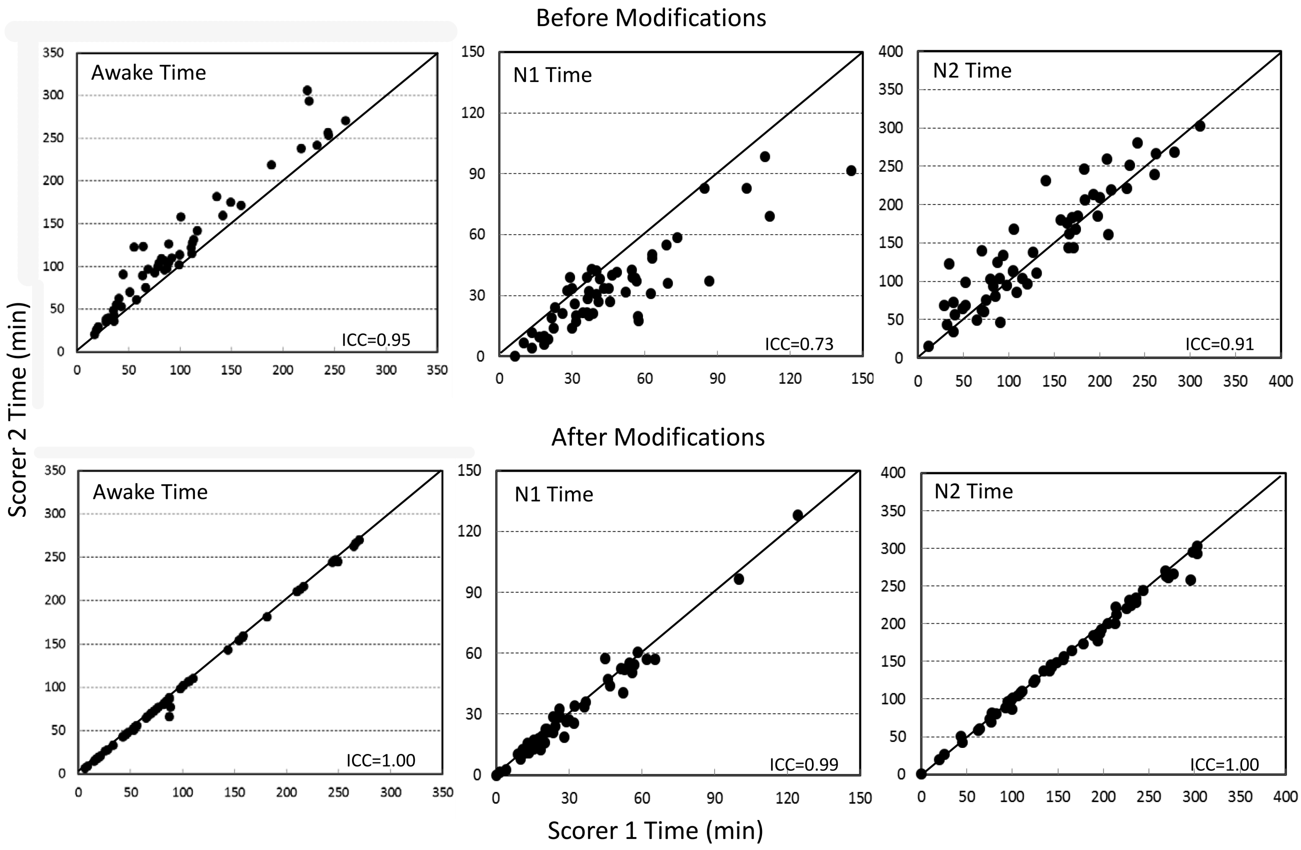
The right half of **Table 1** lists ICCs for the relation between the two scorers and between each scorer's original and modified scores. Before modifications, ICCs for scorer 1 versus scorer 2 were lowest for N1 and N3. Because, unlike Pearson correlation coefficients, ICCs are sensitive to differences in the average of the values being correlated, the reduction in average N1 time for both scorers (left side of **Table 1**) resulted in low ICCs for correlations between premodification and postmodification for this variable (0.45 and 0.67 for scorer 1 and scorer 2, respectively). The modifications appeared to favor scorer 1 with respect to N3 because the ICC for premodification versus postmodification comparison was much higher for scorer 1 (0.93 vs. 0.71). Agreements between premodification and postmodification were excellent and similar for both scorers for stage awake and total sleep time (ICC > 0.9) and very good for N2 (ICC = 0.88). In summary, the modified times agreed well (ICC ≥ 0.88) with the initial times of both scorer 1 and scorer 2 except for N1 time (both scorers) and N3 time of scorer 2.

Errors Resulting from the Proposed Approach

The percent of epochs in an average PSG where the original manual score was not altered by this procedure was 79.5 ± 8.0% for scorer 1 (**Figure S3**, top panel, in the supplemental material) and 79.9 ± 11.2% for scorer 2 (**Figure S3**, bottom panel). Percent of epochs where the stage was changed to that assigned by the other scorer was 8.5 ± 4.6% for scorer 1 and 8.4 ± 6.9% for scorer 2. In the remaining epochs (12.0 ± 6.5% for scorer 1 and 11.7 ± 6.6% for scorer 2) the change was to a different stage (third score). The third score was assigned to the same epoch in one or more of the other five sessions in approximately 40% of cases (4.9 ± 4.2% and 4.8 ± 4.2% of total epochs). Average number of sessions in which the third score was assigned within the other five sessions was 2.0 ± 0.3 for both scorers. The third score was not seen in other sessions in 7.1 ± 3.7% for scorer 1 and 6.9 ± 3.8% for scorer 2 and are considered errors. Most errors involved changing a unanimous awake score to NREM sleep (2.8 ± 3.8% and 3.6 ± 3.8%) and changing a unanimous N1 score to N2 (2.1 ± 1.3% and 1.4 ± 1.1%). In summary, the scores of the two technologists were affected to nearly the same extent and the modified score was independently assigned by one or more technologists in all but ≈ 7% of epochs on average.

Although average number of errors was low for all error categories, there were outliers particularly in the W→NR category (**Figure S4** in the supplemental material). The patient with the most errors (PSG #106) had severe insomnia (sleep efficiency 12%) along with low ORP during stage W. The low ORP during stage W was due to an error in an ORP correction factor that occurs rarely when the PSG contains few sleep and

Figure 1—Agreement between the two scorers in times spent awake and in stages 1 and 2 of non-rapid eye movement sleep.



Top panel: before modifications. Bottom panel: after modifications. Each dot is a separate polysomnogram. ICC = intraclass correlation coefficient.

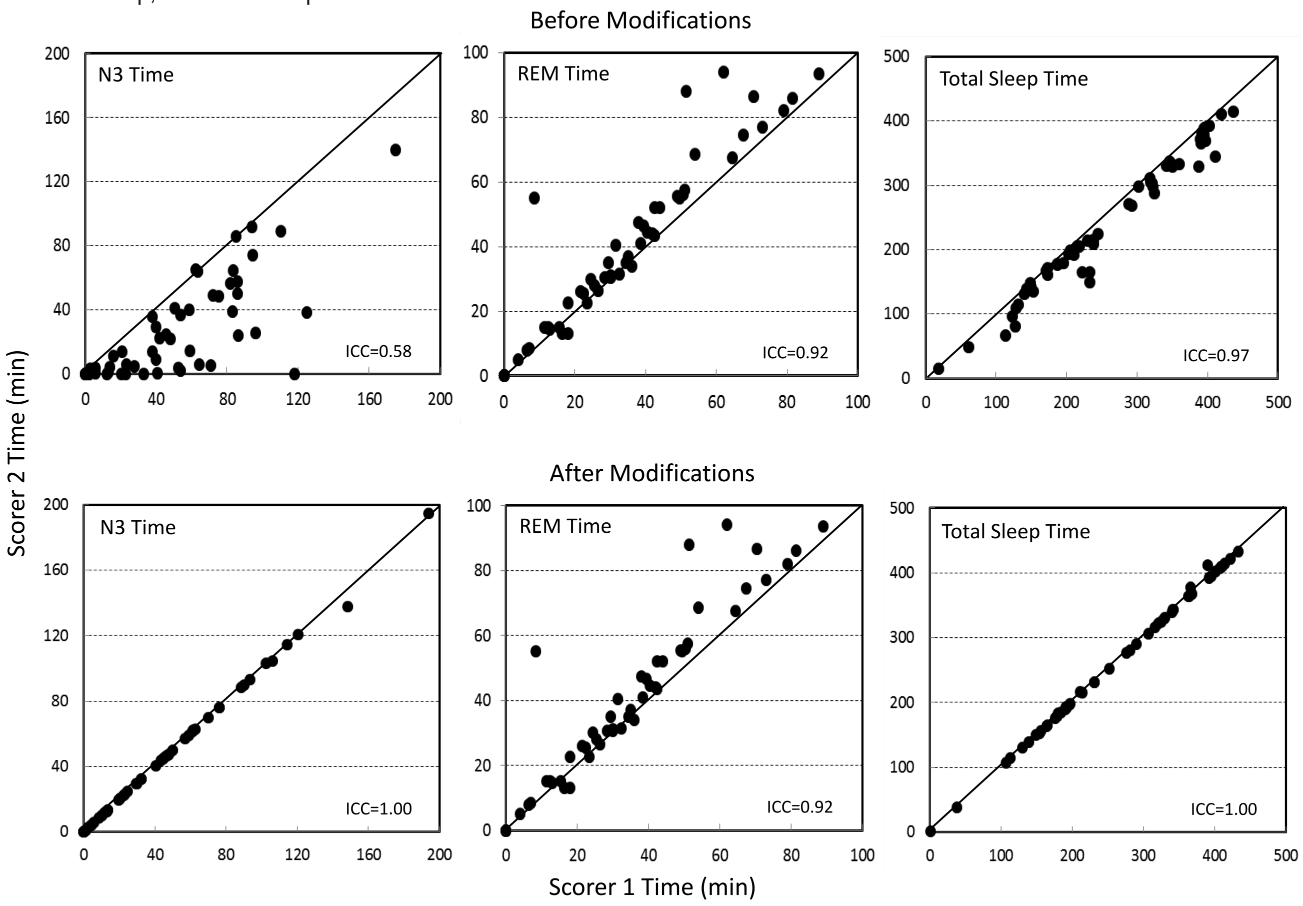
Table 1—Sleep variables before and after scoring modifications.

Variable	Averages (n = 56)				Intraclass Correlation Coefficient			
	S1 Pre	S2 Pre	S1 Post	S2 Post	S1 Pre S2 Pre	S1 Post S2 Post	S1 Pre S1 Post	S2 Pre S2 Post
Awake (min)	107 (85)	127*** (88)	104 (85)	103^{+++} (85)	0.95	1.00	0.93	0.91
N1 (min)	46 (27)	33*** (21)	30^{+++} (23)	29^{+} (23)	0.73	0.99	0.45	0.67
N2 (min)	132 (74)	143* (74)	156^{+++} (81)	153^{+} (79)	0.91	1.00	0.88	0.88
N3 (min)	49 (39)	26*** (31)	40^{+++} (41)	40^{+++} (41)	0.58	1.00	0.93	0.71
REM (min)	32 (23)	37*** (27)	32 (23)	37^{a} (27)	0.92	0.92	1.00	1.00
TST (min)	260 (107)	239*** (106)	259 (110)	259^{+++} (110)	0.97	1.00	0.96	0.95

*,***, significantly different from S1, $p < 0.01$ and $p < 0.0001$, respectively. $^{+,+++}$, significantly different from same scorer before modification, $p < 0.05$ and $p < 0.0001$, respectively. ^{a}, significantly different from S1 post, $p < 0.0001$. S1, scorer 1; S2, scorer 2; pre, before modification; post, after modification; N1, N2, N3, non-rapid eye movement stages 1, 2 and 3; REM, rapid eye movement; TST, total sleep time.

many awake epochs with low beta power and frequent REM. In another patient, electrocardiogram artifacts in the EEG were too wide (0.18 sec) to be detected/removed by the R-wave removal algorithm because of left bundle branch block, thereby artificially elevating theta power, which reduces ORP.[16] In two patients the EEG in the affected sections was visually indistinguishable from stage N1 but there were behavioral indications of wakefulness (REM with high chin EMG). Finally, in one patient what visibly appeared to be alpha waves during wakefulness had a dominant frequency in the theta range (5.7 Hz),

Figure 2—Agreement between the two scorers in times spent in rapid eye movement (REM) sleep, stage 3 of nonrapid eye movement sleep, and total sleep time.



Top panel: before modifications. Bottom panel: after modifications. Each dot is a separate polysomnogram. The lack of change in REM times is because epochs scored as REM were not modified as per protocol. ICC = intraclass correlation coefficient.

thereby artificially increasing theta power and reducing ORP during stage W. True errors occurred in less than 10% in all but the same five patients (**Figure S4**).

Consistency in Scoring Different Sleep Stages

Figure 3 shows the likelihood of agreement among the seven scoring sessions when a sleep stage was scored by at least one scorer in a single scoring session. For stage W, there were 14,956 epochs in which stage W was seen at least once. In 64.9% of these the score was unanimous (7/7 sessions). The percent varied widely among PSGs (43.9–87.9%), indicating that in some PSGs scoring of this stage was more challenging than in others. The likelihood of the stage not being scored in any other session (1/7) was very small (6.6 ± 4.8%) and a plurality (4/7) was found in 80.8 ± 11.7% (64.7–95.0%) of these epochs. A similar pattern was seen for REM sleep with unanimity found in 62.7 ± 21.9% (CI: 33.1–87.9%) and a plurality in 82.5 ± 13.1% (CI: 65.5–93.6%). By contrast, for stage N1, unanimity was found in only 9.7 ± 6.5% of epochs (1.4–18.5%) and a plurality was found in only 37.2 ± 12.0% staged at least once as N1 (21.3–52.8%). Stage N3 showed a pattern similar to that of stage N1 whereas the pattern for stage N2 was intermediate with unanimity seen in 46.0 ± 16.5% (25.7–66.4%) and a plurality seen in 72.8 ± 13.1% (56.7–85.7%).

DISCUSSION

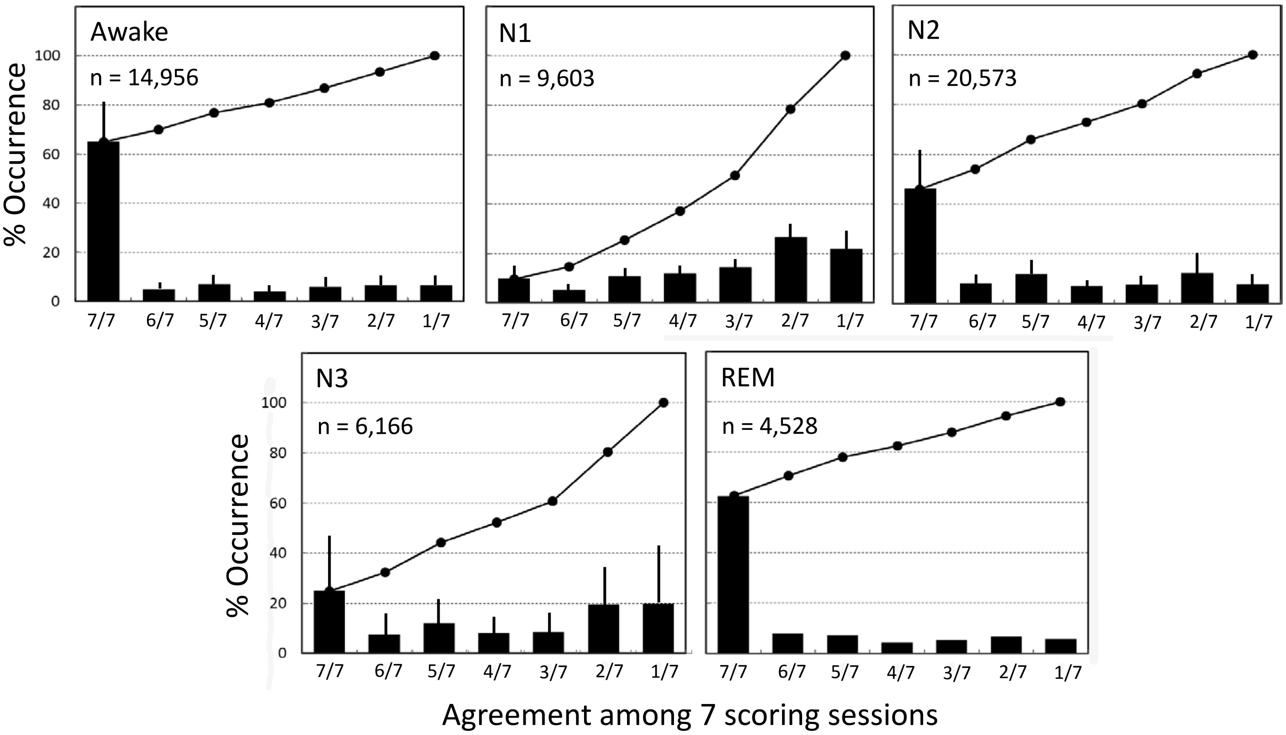
The current study indicates that providing technologists with digitally obtained information regarding ORP, spindles, K complexes, and delta wave duration during scoring can greatly assist them in scoring epochs with equivocal features and dramatically reduce interrater variability.

Although inconsistency among scorers for stages N1 and N3 is well known, our current finding that only a small fraction of epochs scored as N1 or N3 by one scorer in one scoring session are scored the same by all or even by a simple majority of competent scorers (**Figure 3**) indicates that scoring these stages is little more than a guessing game. This further emphasizes the need to address the issue of interrater variability.[11]

Use of Digitally Obtained EEG Features as an Adjunct to Manual Sleep Scoring

As mentioned earlier, most differences between scorers occur in epochs with equivocal EEG features[12] and the types of

Figure 3—Extent of agreement among the seven scoring sessions when a given sleep stage is scored at least once in one of the sessions.



W, stage awake; REM, rapid eye movement sleep; N1, N2, and N3 are nonrapid eye movement stages 1, 2 and 3. Note that when stage awake occurred in one scoring session there was unanimous agreement (7/7) about the stage in 65 % of cases. A similar behavior was seen in stage REM. However, when either N1 or N3 was scored there was unanimous agreement in only a small fraction of the epochs, whereas in more than half the cases the score was seen in only one or two other sessions. Diagonal lines are the cumulative occurrences. Vertical bars are standard deviation. n, number of epochs in each category.

scoring differences suggest that the equivocal features relate to three issues: (1) whether in an epoch with feature of both wakefulness and sleep, awake features occupy > 15 sec, (2) whether spindles or K complexes are present, and (3) whether delta wave duration exceeds 6 sec in a 30-sec epoch. This study is the first to use digitally acquired values for these three features as an arbitrator in such cases. The results show that doing so does in fact reduce interrater variability to a negligible level; % agreement increased from 77.6 ± 10% in individual PSGs to 96.5 ± 2.6% (**Figure S2**). At the same time, with very few exceptions, true errors resulting from this arbitration process were within an acceptable range (≈ 7%, **Figures S3** and **S4**). Furthermore, agreement between the modified sleep score and the original manual score of each technologist (last two columns, **Table 1**) was well within reported agreement between two competent scorers for all sleep stages.[1–11] Thus, even without human oversight (**Table 1**), the modified score is equivalent to one produced by a competent third scorer whose scores agree well with other competent scorers. However, it differs from any other single scorer in that, if used exclusively, interrater variability in sleep staging would be significantly reduced.

Distinction between Awake and NREM Sleep

We used an ORP of 1.5 as a cutoff between wakefulness and sleep. In development and validation tests (internal validation and published studies),[16,26] we found that 30-sec epochs with an average ORP > 2.0 are consistently scored awake in > 95% of cases, whereas epochs with average ORP < 1.0 are consistently scored asleep. Epochs with ORP between 1.0 and 2.0 contain both awake and sleep features. Sleep onset is identified almost invariably within this range[27] and most W/NREM discrepancies occur in epochs with ORP between 1.0 to 2.0.[16] However, there is no absolute ORP cutoff within this range that distinguishes with certainty epochs that are manually scored as awake from those scored as sleep. Thus, choosing a unique level to arbitrate the stage will inevitably introduce errors in that some epochs can be confidently scored as sleep when ORP is higher than the cutoff level, and *vice versa*. A value of 1.5 was our first choice simply because it is in the middle of the transitional range. Too high a cutoff would result in too many epochs converted to sleep when they are unequivocally awake, and *vice versa*. A perfect compromise would be a level in which the number of unequivocally awake epochs that are converted to sleep equals the number of unequivocally asleep epochs that are converted to awake and both numbers are acceptably small.

The results show that with a threshold of 1.5 the number of epochs converted to sleep against a unanimous decision was ≈ 3% of total epochs, whereas the opposite occurred in only 1.0 % of epochs (**Figure S3**). This suggests that 1.5 was perhaps a little too high. However, we elected not to experiment with other thresholds in view of the small number of errors involved, particularly because the epochs subjected to this error are already transitional between awake and asleep and do not reflect full wakefulness. There were only a few PSGs (5 of 56) in which an excessive number of epochs (> 8%) were erroneously converted to sleep (**Figure S4**). These errors, however, would not have been averted by a small reduction in the cutoff value because they were related to unusual EEG patterns and artifacts and the ORP in the affected epochs was well below 1.5.

Distinction between NREM Stages N1 and N2

Manual scoring of spindles and K complexes is subject to much variability and uncertainty.[14,15] It was estimated that "2 to 3 experts are needed to build a spindle scoring dataset with substantial reliability and 4 or more experts are needed to build a dataset with near perfect reliability".[15] A secondary contributing factor is interrater variability in scoring arousals because the stage is changed to N1 following arousals. Thus, discrepancies in arousal scoring result in N1/N2 discrepancies unless spindle/K complex frequency is sufficiently high that conversion to N1 following arousal is immediately reversed.

A number of signal processing techniques have been used to develop several automated methods for detection of sleep spindles and K complexes.[17–25] We elected to use the algorithms built into the commercial MSS system for a number of reasons: (1) they were available to us, (2) if shown to be effective they can be implemented immediately (the system is available commercially), (3) the algorithms were developed by trial and error with the specific aim of achieving the best agreement between highly competent scorers, which is the main aim of the proposed approach, (4) agreement between MSS, which utilizes these algorithms, and the average of 10 academic scorers in scoring N1 and N2 was shown to be superior to agreement across academic sites and comparable or superior to agreement between scorers at the same site,[9] and (5) none of the methods described in the literature has been generally adopted because of its superiority to others or its acceptable performance against a consensus of several scorers.[15]

It is predictable that if decisions between N1 and N2 are made based on detection of spindles/K complexes by a single independent scorer (automatic system in this case), most N1/N2 discrepancies would disappear regardless of how accurate the automatic system is. This was clearly confirmed here; N1/N2 disagreements decreased from an average 42 ± 36 epochs/PSG to 10 ± 10 epochs/PSG. The remaining disagreements were the result of discrepancies in arousal scoring. Accordingly, the main question to be addressed is the accuracy of the algorithms used here to detect these events. Should the algorithms be inaccurate, the benefit of minimizing N1/N2 disagreements would clearly be offset by a large number of epochs that would erroneously be converted from N1 to N2, and *vice versa*. Evidence that the algorithms performed adequately for the intended purpose was obtained from two sets of observations. First, frequency of spindles and K complexes was highest in stage N2 (**Figure S1**) and when they were detected in stage N1 they were consistent with the manual stage assigned by the scorers except in a very few cases (43 spindles, and no K complexes, in the entire dataset). Spindles were falsely identified during epochs staged awake (**Figure S1**) and during arousals. However, as long as spindles identified under these conditions are ignored, as was done here, there should be very few errors. Second, only 2.1 ± 1.3% of epochs in the case of scorer 1 and 1.4 ± 1.1% in the case of scorer 2 were changed from N1 to N2 in the face of a unanimous N1 decision (**Figure S3**) and the largest such error in any PSG was 6% (**Figure S4**). Likewise, there were very few epochs that were changed from N2 to N1 in the face of a unanimous N2 score (0.5 ± 0.5 and 0.3 ± 0.4% for scorer 1 and scorer 2, respectively; **Figure S3**), with the largest error being 2.5% (**Figure S4**). We believe this is an acceptable compromise for a marked improvement in inter-scorer agreement.

Distinction between NREM Stages N2 and N3

N2/N3 discrepancies arise from difficulty in visually estimating total delta wave duration in borderline epochs. To do that visually requires identification of each wave that meets the amplitude (75 μV) and duration (0.5–2.0 sec) criteria, measuring the duration of each eligible wave and summing all durations. This is clearly not practical and scorers simply eyeball the epoch. As in the case of spindles and K complexes, if an estimate of delta duration made by a single independent scorer (the automatic system in this case) is enforced, perfect agreement would result, regardless of the estimate's accuracy. As expected, the proposed approach reduced N2/N3 disagreements from 2,849 epochs (52 ± 47 epochs/PSG) to 22 epochs (< 1 epoch/PSG) (**Figure S2**). The question is whether the algorithm is accurate enough such that very few epochs unequivocally scored as N2 are converted to N3, and *vice versa*. Such errors were very few (0.5 ± 0.7 epochs per PSG for erroneous N2 to N3 conversion and 0.2 ± 0.3 epochs per PSG for the opposite error; **Figure S3** and **S4**), thereby indicating that the algorithm is adequate for this purpose.

How to Implement the Proposed Approach

Most PSG acquisition systems offer scoring modules to assist in the manual scoring of various PSG variables. One or more of the three algorithms used here can be added to the bank of modules. With every epoch exposed on the viewer these modules would display average ORP, average delta wave duration and/or whether spindles and K complexes are present in the first or second half of the epoch. As shown before,[12] scoring differences among competent technologists develop primarily when epochs are ambiguous such that one scorer may be leaning toward a score while the other, or the same scorer at another time, may lean away from it. If technologists are instructed to be guided by the results of the modules unless they are absolutely certain the modules are wrong, the guessing aspect of scoring such epochs is removed, resulting in much better agreement between scorers. Such an approach may also reduce scoring time because the interpretation of ambiguous epochs can be challenging and time consuming.

A special caveat must be pointed out regarding wake/sleep decisions. Here, the use of 15 sec of "awake" time as cutoff between wakefulness and sleep[13] would have to be abandoned. Epochs in which this rule applies are *ipso facto* transitional. Determination of whether the alpha and/or beta pattern span exactly 15 sec is often difficult and is the main reason for wake/

sleep discrepancies. The use of an objective measure of sleep depth (e.g., ORP) in such epochs is preferred because it reflects the complex features of the entire epoch[16] and not a single parameter such as duration of a specific pattern. Furthermore, we have shown that, with few exceptions, implementing an ORP cutoff of 1.5, regardless of what the technologist might score using the 15-sec rule, is associated with a clinically insignificant number of epochs where the "modified" score is not assigned by one or more scorers in other sessions (**Figure S4**). Accordingly, in using the ORP module, technologists should be advised to score epochs as asleep if ORP is < 1.5 even if they are certain that the alpha/beta pattern is > 15 sec unless there are clear signs of wakefulness, such as eye blinks or REMs with high EMG.

Limitations

The proposed approach does not address REM/NREM scoring discrepancies. We elected not to address this issue at this point because there was little disagreement in REM scoring in this data set (**Figure 2**); consequently, no significant benefit can reasonably be expected.

The very small number of unjustified spindles found in epochs with unanimous N1 score (43 spindles in 55 PSGs) suggests that the algorithm produces very few false-positive results when used outside arousals and awake periods. As such, the algorithm proved adequate for the purpose of distinguishing stages N1 and N2. However, its accuracy in identifying individual spindles during NREM sleep was not tested. Accordingly, it currently cannot be recommended for use to evaluate frequency of spindles or spindle characteristics.

Sampling frequency of the EEG/electro-oculography/EMG signals was lower (128 Hz) than that currently recommended by the American Academy of Sleep Medicine (AASM) (≥ 200 Hz). However, the higher frequency recommended by AASM is simply to improve the visual resolution during manual scoring. It has no effect on digital scoring or ORP and other features because the original EDF is downsampled to a standard 120 Hz and filtered according to AASM criteria before it is digitally processed.

We only analyzed PSGs recorded with one acquisition system (Sandman). However, there is no technical reason to expect that the proposed approach would not work with other systems as long as the system follows the AASM recommendations for data acquisition and the acquired data can be exported in the EDF format without prior filtering. Sampling frequency is irrelevant provided it exceeds 120 Hz because the original EDF is first downsampled to a standard 120 Hz by MSS before it is subjected to digital analysis.

CONCLUSIONS

This study has shown that currently available digital methods that determine depth of sleep and delta wave duration and identify sleep spindles and K complexes have sufficient accuracy to guide the scoring of difficult-to-score epochs. Making them available during scoring can result in dramatic reduction in interrater variability in scoring sleep stages.

ABBREVIATIONS

AHI, apnea-hypopnea index
CPAP, continuous positive airway pressure
EEG, electroencephalogram
EMG, electromyogram
ICC, intraclass correlation coefficient
MSS, Michele sleep scoring system
N1, stage 1 of non-rapid eye movement sleep
N2, stage 2 of non-rapid eye movement sleep
N3, stage 3 of non-rapid eye movement sleep
NREM, non-rapid eye movement sleep
ORP, odds ratio product
OSA, obstructive sleep apnea
PLM, periodic limb movement
PSG, polysomnogram
REM, rapid eye movement sleep
TST, total sleep time
W, stage awake

REFERENCES

1. 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.
2. 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.
3. Collop NA. Scoring variability between polysomnography technologists in different sleep laboratories. Sleep Med 2002;3:43–7.
4. Danker-Hopfe H, Kunz D, Gruber G, et al. Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders. J Sleep Res 2004;13:63–9.
5. Pittman SD, MacDonald MM, Fogel RB, et al. Assessment of automated scoring of polysomnographic recordings in a population with suspected sleep-disordered breathing. Sleep 2004;27:1394–403.
6. Anderer P, Gruber G, Parapatics S, et al. An E-health solution for automatic sleep classification according to Rechtschaffen and Kales: validation study of the Somnolyzer 24 × 7 utilizing the Siesta database. Neuropsychobiology 2005;51:115–33.
7. Magalang UJ, Chen NH, Cistulli PA, et al. Agreement in the scoring of respiratory events and sleep among international sleep centers. Sleep 2013;36:591–6.
8. Kuna ST, Benca R, Kushida CA, et al. Agreement in computer-assisted manual scoring of polysomnograms across sleep centers. Sleep 2013;36:583–9.
9. 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.
10. Zhang X, Dong X, Kantelhardt JW, et al. Process and outcome for international reliability in sleep scoring. Sleep Breath 2015;19:191–5.
11. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 2013;9:81–7.
12. Younes M, Raneri J, Hanly P. Staging sleep in polysomnograms: analysis of inter-scorer variability. J Clin Sleep Med 2016;12:885–94.
13. Berry RB, Brooks R, Gemaldo CE, Harding SM, Marcus CL, Vaughn BV for the American Academy of Sleep Medicine. The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications, version 2.0. www.aasmnet.org. Darian, IL: American Academy of Sleep Medicine, 2012.
14. Warby SC, Wendt SL, Welinder P, et al. Sleep-spindle detection: crowdsourcing and evaluating performance of experts, non-experts and automated methods. Nat Meth 2014;11:385–92.

15. Wendt SL, Welinder P, Sorensen HB, et al. Inter-expert and intra-expert reliability in sleep spindle scoring. Clin Neurophysiol 2015;126:1548–56.

16. Younes M, Ostrowski M, Soiferman M, et al. Odds ratio product of sleep EEG as a continuous measure of sleep state. Sleep 2015;38:641–54.

17. Martin N, Lafortune N, Godbout J, et al. Topography of age-related changes in sleep spindles. Neurobiol Aging 2013;34:468–76.

18. Ferrarelli F, Huber R, Peterson MJ, et al. Reduced sleep spindle activity in schizophrenia patients. Am J Psychiatry 2007;164:483–92.

19. Wamsley EJ, Tucker MA, Shinn AK, et al. Reduced sleep spindles and spindle coherence in schizophrenia: mechanisms of impaired memory consolidation? Biol Psychiatry 2012;71:154–61.

20. Mölle M, Marshall L, Gais S, Born J. Grouping of spindle activity during slow oscillations in human non-rapid eye movement sleep. J Neurosci 2002;22:10941–7.

21. Bódizs R, Körmendi J, Rigó P, Lázár AS. The individual adjustment method of sleep spindle analysis: methodological improvements and roots in the fingerprint paradigm. J Neurosci Meth 2009;178:205–13.

22. Wendt SL, Christensen JA, Kempfner J, et al. Validation of a novel automatic sleep spindle detector with high performance during sleep in middle aged subjects. Conf Proc IEEE Eng Med Biol Soc 2012;2012:4250–3.

23. Devuyst S, Dutoit T, Stenuit P, Kerkhofs M. Automatic K-complexes detection in sleep EEG recordings using likelihood thresholds. Conf Proc IEEE Eng Med Biol Soc 2010;2010:4658–61.

24. Lajnef T, Chaibi S, Eichenlaub JB, et al. Sleep spindle and K-complex detection using tunable Q-factor wavelet transform and morphological component analysis. Front Hum Neurosci 2015;9:414.

25. Ray LB, Sockeel S, Soon M, et al. Expert and crowd-sourced validation of an individualized sleep spindle detection method employing complex demodulation and individualized normalization. Front Hum Neurosci 2015;9:507.

26. Younes M, Thompson W, Leslie C, Egan T, Giannouli E. Utility of technologist editing of polysomnography scoring performed by a validated automatic system. Ann Am Thorac Soc 2015;12:1206–18.

27. Meza S, Giannouli E, Younes M. Enhancements to the multiple sleep latency test. Nat Sci Sleep 2016;8:145–58.

ACKNOWLEDGMENTS

The authors thank Colleen Leslie, John Laprairie, and Michele Ostrowski for scoring the PSGs and for coordinating the study.

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication April, 2016
Submitted in final revised form June, 2016
Accepted for publication June, 2016

Address correspondence to: Magdy Younes, MD, 1001 Wellington Crescent, Winnipeg, Manitoba Canada R3M 0A7; Tel: (204) 297-0978; Fax: (866) 693-5084; Email: mkyounes@shaw.ca

DISCLOSURE STATEMENT

Support was provided by YRT Ltd, Winnipeg, Manitoba, Canada. Dr. Younes is majority owner of YRT Ltd. Dr. Hanly has indicated no financial conflicts of interest.