Expanded Methods

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 Odds-Ratio-Product[1] and to determine the reasons for inter-rater variability.[2] 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 oro-nasal thermister, oxyhemoglobin saturation, and a microphone. They were recorded with a Sandman system using SD 32 global amplifiers (Natus Medical, Pleasanton, CA) with no hardware filters applied. Digital resolution was 14 bits/sample. Sampling rate was 128 Hz for all non-respiratory channels. When manually scored, digital filter settings for the EEG and EOG were 0.3 and 35.0 Hz for the high and low pass filters, respectively. A 10Hz high pass filter with no low pass filter was applied to the EMG signal. 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 years 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[3,4] automatic system (Michele Sleep Scoring, MSS, Winnipeg, Manitoba, Canada). Scorer 1, scorer 2 and a third senior scorer (scorer 3) from Calgary were asked to edit the automatic score, epoch by epoch, and correct any score they disagreed with. Accordingly, there were 7 manual scores for each epoch in each PSG, 3 each from scorer 1 and scorer 2 and one from scorer 3.

To perform Auto-scoring, the EDF file is uploaded in the MSS software (including viewer). A new EDF is generated with the characteristics required by the scoring algorithm. These include down-sampling the signals to 120 Hz regardless of the original sampling. The scoring software then applies the recommended filters (0.3 Hz high pass and 35 Hz low pass filters) to the EEG and EOG signals and 10Hz high pass filter to the EMG signal.

The method of calculating ORP was described in detail elsewhere.[1] Briefly, fast Fourier transform is applied to the EEG (C3 and C4) in 3-second non-overlapping sections to generate the powers in the frequency range 0.33 Hz to 60 Hz in 0.33 Hz increments. For each 3-second epoch the sum of powers in each of the following ranges is calculated: 0.33-2.33 Hz (delta), 2.37-6.33 Hz (theta), 7.33 to 14.00 Hz (alpha/sigma), and 14.33-35.00 Hz (beta). Powers at 6.67 and 7.00 Hz, normally included in the theta range, are not included in any category since such waves can exist during full wakefulness in some patients.[1] The database of MSS contains power values in each of these four ranges collected from >400,000 epochs obtained PSGs with assorted pathology. The entire range of power values in each frequency band in the database is divided into 10 equal aliquots which are ranked 0 to 9 (9 contains the highest 10% of the range). The

EEG within each 3s epoch is assigned one of 10,000 patterns based on the relative powers in the four frequency ranges. For example, pattern 2741 would indicate little delta activity (2), moderately high theta activity (7), average alpha activity (4) and very low beta activity (1). The database also contains a table that indicates the likelihood of each of these patterns occurring in epochs staged awake by a consensus of highly experienced technologists. The likelihood ranges from 100% (always seen during epochs staged awake) to zero (never seen during epochs staged awake. The ORP value is the normalized probability (2.5=100% and 0 = 0%) of the assigned pattern to occur during wakefulness.[1]

The algorithms for scoring spindles and K complexes in MSS were refined over several years by iteratively adjusting the parameters to achieve the highest possible agreement with manual scoring of N1 and N2 by expert technologists in hundreds of PSGs. For spindles, spectral analysis (fast Fourier transform) is performed on overlapping 1s epochs with 0.2s difference between successive 1s epochs, giving a resolution of 0.2s (5 values per second). For each one second of data the power in each frequency between 1 and 60 Hz is obtained and the sum of powers in the frequency range 10-16 Hz is calculated (Power S). For each 30-second epoch the 30^{th} percentile of the 150 power S values is identified. A running ratio (Ratio-S) of current power S to the 30^{th} percentile of power S in the 30s epoch is obtained. A spindle is identified when Ratio-S exceeds 3 or (current Power S – 30^{th} percentile) >12 μV^2 for > 5 but <13 consecutive measurements. This allows for spindles to range in duration from 0.4 to 2.2s.

For K complexes a 3Hz low pass filter is applied to the EEG signal. A K complex is identified when there is a negative deflection in the filtered EEG with an amplitude >30μV, a most negative value >4*average of all negative values in the preceding 4s, a time to peak negative value between 0.10 and 0.45s, no positive wave >25μV in the preceding 1.5s, a positive wave

following the initial negative deflection that is > the pre-complex most positive wave, and a total duration between 0.5 and 2.5s.

For delta wave duration, C3 and C4 EEG signals are low-pass filtered (3Hz) and waves > 75μV in peak to peak amplitude and duration (negative zero crossing to next negative zero crossing) of 0.4 to 2.0 sec are identified. The sum of delta wave durations in 30s epochs is calculated for each derivation and the two sums are averaged.

Following auto-scoring, Excel files were generated that listed average ORP and average delta duration in each 30s epoch and the location of each spindle and K complex identified by the program. 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 seconds 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 was 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 the change was carried forward until the manual stage was no longer N1 or 4 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* in the main manuscript).

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

Epochs staged manually as REM sleep were not modified.

REFERENCES

1. 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.
2. Younes M, Raneri J, Hanly P. Staging Sleep in Polysomnograms: Analysis of Inter-Scorer Variability. J Clin Sleep Med 2016;12:885-94.
3. 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.
4. 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.

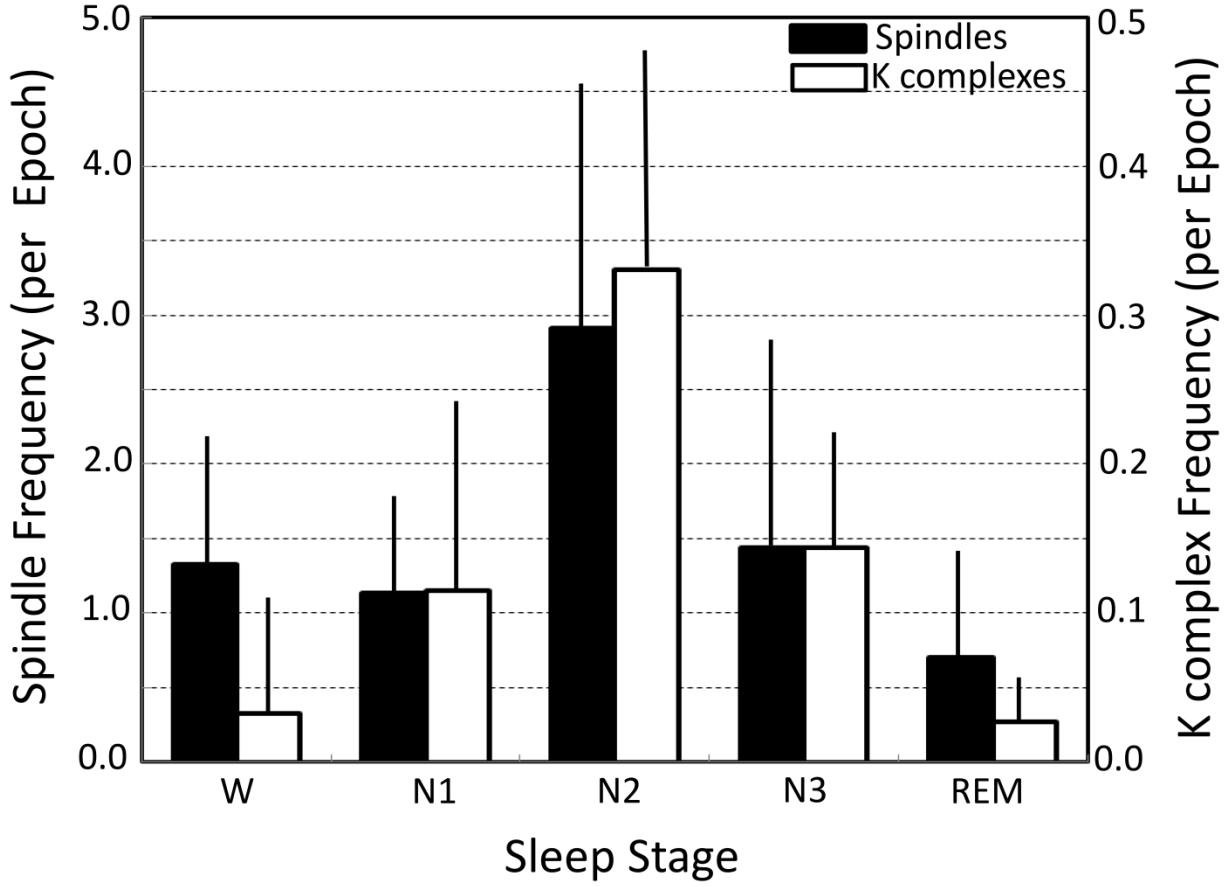

Figure S1—Frequency of digitally-identified sleep spindles and K complexes in different sleep stages. The epochs used for this analysis are those in which sleep stage was the same in all scoring sessions (i.e. unanimously scored). W, stage awake; REM, rapid eye movement sleep; N1, N2, and N3 are non-REM stages 1, 2 and 3. Note the different scales for spindles (left) and K complexes.

A

Before Modification

S1	S2				
	W	N1	N2	N3	REM
W	11561	81	61	1	26
N1	1979	2022	787	26	253
N2	385	1493	12077	145	397
N3	8	28	2703	2660	2
REM	46	19	56	1	3443

% Agreement = 78.9% kappa = 71.1%

B

After Modification

S1	S2				
	W	N1	N2	N3	REM
W	11688	0	2	0	108
N1	0	2834	260	4	152
N2	2	288	16486	5	414
N3	5	10	17	4402	4
REM	21	29	67	5	3443

% Agreement = 96.5% kappa = 95.1%

Figure S2—Epoch by epoch comparison of 5-stage sleep scoring between two scorers (scorer 1 [S1] and scorer 2 [S2]) before and after the modifications of the proposed approach. W, stage awake; REM, rapid eye movement sleep; N1, N2, and N3 are non-REM stages 1, 2 and 3.

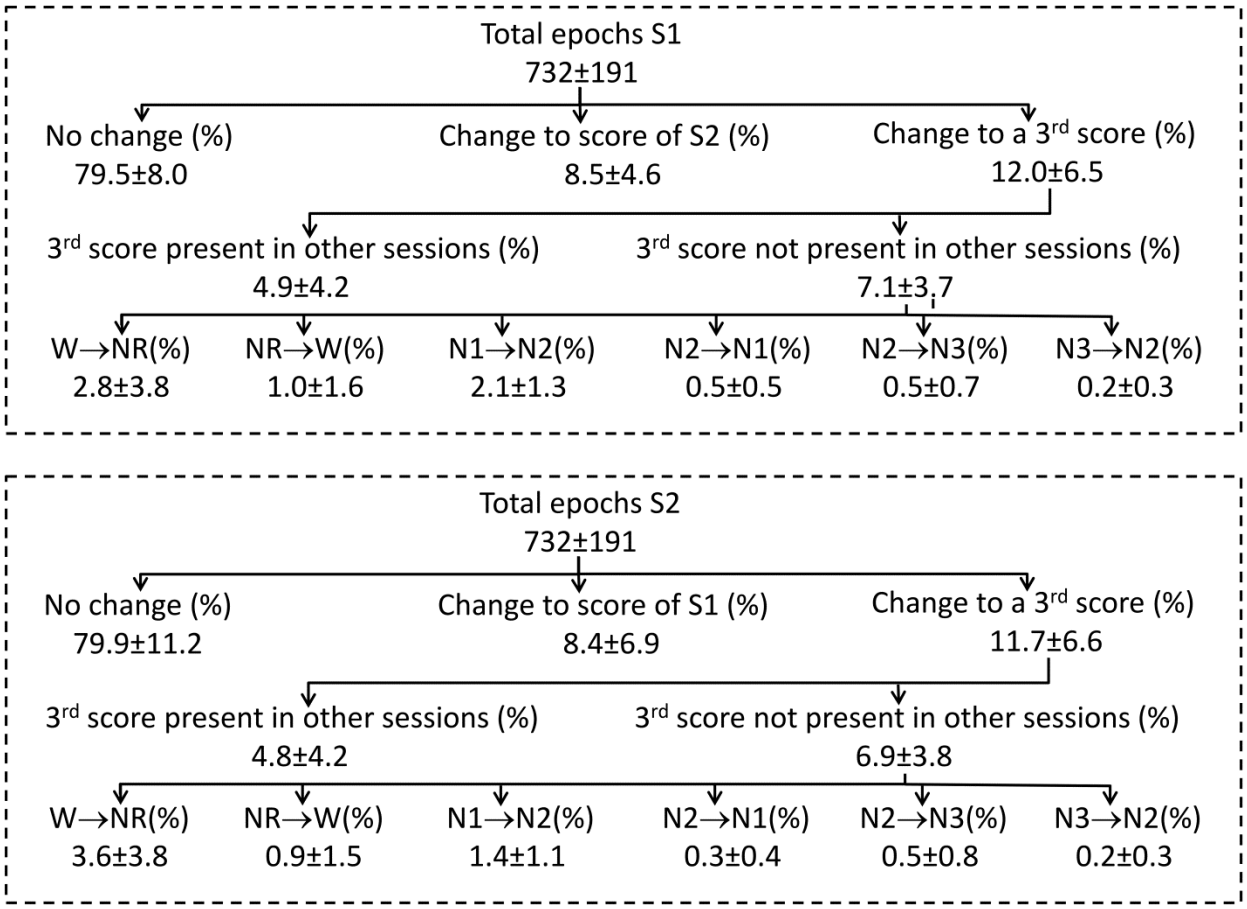

Figure S3—Diagram showing the % of epochs ± SD that were not affected by the procedure (No change), % in which the score of one technologist was independently changed to that of the other, and % of epochs in which the modified score was different from both scorers (scorer 1 [S1] and scorer 2 [S2]). Epochs in the last category are divided into those in which the different score assigned by the proposed approach was assigned during one or more of the other manual sessions, and epochs where the different score was not seen in any other session (errors). Errors were classified into 6 categories. W, stage awake; NR, non-rapid-eye-movement sleep; N1, N2, and N3 are NR stages 1, 2 and 3. Results for S1 and S2 are shown in the top and bottom panels, respectively. Note that the modifications to the scores of each technologist were done without knowledge of the other technologist's score.

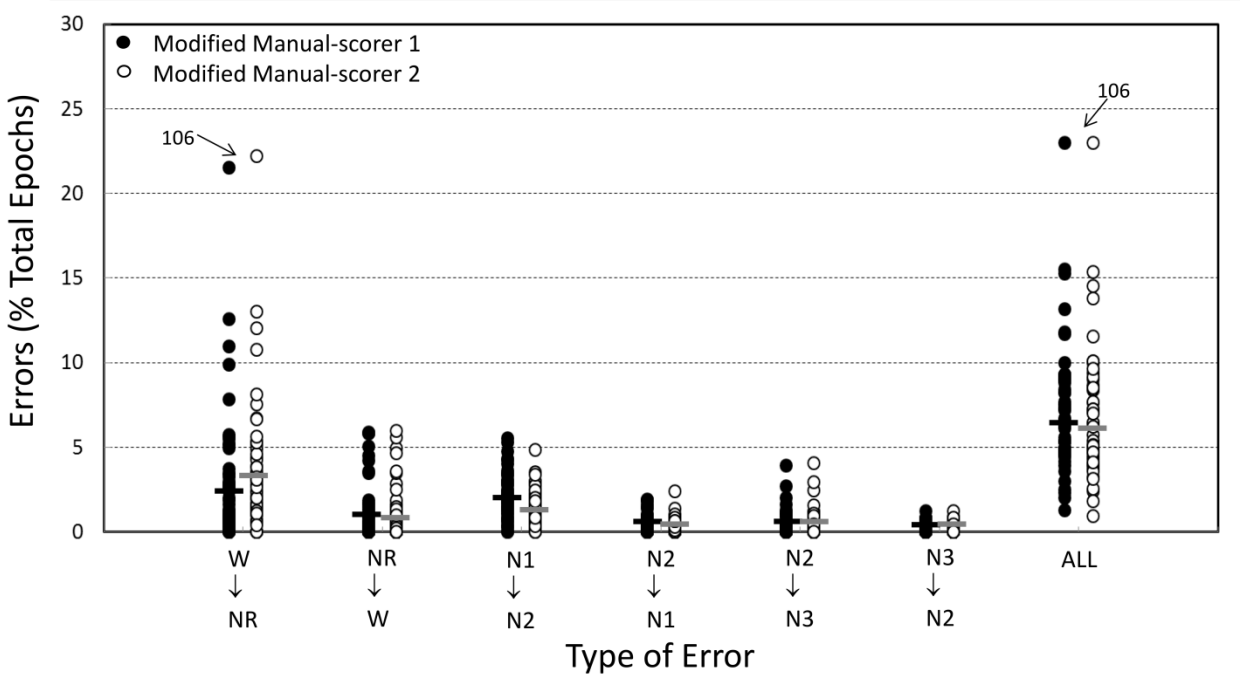

Figure S4—Percent of epochs with different errors in individual polysomnograms (PSGs). W, stage awake; NR, non-rapid-eye-movement sleep; N1, N2, and N3 are NR stages 1, 2 and 3. All, is the sum of all errors in each PSG.