michele.
SLEEP SCORING

MICHELE SLEEP SCORING SYSTEM

Summary of Features and Performance

michele.
SLEEP SCORING

info@michelesleepscoring.com

1470 Willson Place Unit B / Winnipeg, Manitoba / R3T 3N9
Canada
phone 1 877 949 3202 / fax 866 693 5084

Table of Contents

This report was generated by:

MICHELE Sleep Scoring System

Younes Medical Technologies (YMT)
1470 Willson Place Unit B
Winnipeg, Manitoba R3T 3N9
Canada
Tel: 1-888-942-6774

https://michelesleepscoring.com/

Advena Ltd.
Tower Business Center
2nd Flr, Tower Street
Swatar, BKR 4013
Malta

Intended Use:

The MICHELE Sleep Scoring System is a computer program (software) intended for use as an aid for the diagnosis of sleep and respiratory related sleep disorders.
The MICHELE Sleep Scoring System is intended to be used for analysis (automatic scoring and manual rescoring), display, redisplay (retrieve), summarizing, reports generation and networking of digital data collected by monitoring devices typically used to evaluate sleep and respiratory related sleep disorders.
The device is to be used under the supervision of a physician. Use is restricted to files obtained from adult patients.

Warnings and Cautions:

The MICHELE Sleep Scoring System does not analyze data that are different from those analyzed by human scorers.
The attached report contains numerical and graphical data typical of those generated following manual scoring of polysomnograms. It does not contain any interpretation, diagnosis or recommendations for treatment. These decisions are to be made by the treating physician.
The scoring provided by the MICHELE Sleep Scoring System must be reviewed, edited as necessary and approved before the report is generated.
Special attention should be paid to confirm the first epoch of Rem sleep and of non-Rem sleep since, occasionally, brief periods of either kind can be missed or mistakenly scored resulting in large errors in sleep latency or Rem latency.
Arousals may be underreported in cases where arousal frequency is very high.
Any modification to any component of the MICHELE Sleep Scoring System may result in erroneous results or failure of the software to operate.
Caution: US Federal law restricts this device to sale by or on the order of a physician.

Identifier of the device:

Viewer Software Version:
Release Date:
Autoscoring Software Version:
Release Date:

CE
2797

Introduction

The Michele Sleep Scoring System (Michele) is a software system that automatically scans physiological data obtained during level 1, 2 and 3 sleep studies, referred to as polysomnography (PSG) records, and applies a variety of analytical approaches to identify the occurrence of certain events that relate to the presence and type of sleep state, breathing abnormalities and limb movements. The system scores Sleep Stages, Arousals, Respiratory Events and Leg Movements. At the end of the analysis the system generates a PSG Report that includes tables and graphs typical of those generated following manual scoring of PSG records by certified technologists. The results of the automated scoring may be displayed using the PSG Scoring Viewer (referred to as Viewer in this Summary) application, which allows manual editing of the results and generation of a revised PSG Report.

How the Service is Provided

Users are provided with the Viewer software. The Viewer can be installed in any number of computers in the client laboratory. The digital raw data are exported from the data acquisition system in EDF format. The EDF file is then opened in the Viewer. The user answers several questions that describe the file characteristics (e.g. channel mapping, whether the signals are pre-filtered...etc), what scoring options are desired (e.g. criteria for scoring hypopneas), and what information should go into the report. If the montage is always the same and the options do not change, this procedure need only be done once with the first file. Once these selections are confirmed, the user uploads the file to a secure server by clicking on a "Score File" button. The speed of the uploading process varies depending on the bandwidth provided by the client's Internet provider, but typically takes in the order of 5 minutes. Once the file is uploaded, Michele detects the new file and scores it. Scoring is typically complete in 4 to 6 minutes, depending on file size. Once scoring is complete, the Viewer indicates that the scoring results are ready. The results are downloaded in less than a minute by clicking on the Download button. The automatic scoring can then be edited by the physician or technologist, who may also refer to the Editing Helper function built into the Viewer, which scans the study and recommends sections to be reviewed. Once the scoring is approved, a **fully customizable report** can be generated to be viewed by interpreting physicians.

Features of the viewer are very similar to those of other Viewers (e.g. Sandman, Alice) and, additionally, include other proprietary features such as computation of breath amplitude (% baseline), total duration of Delta waves in the epoch, and presence of flow limitation. Other unique information, e.g. an index of sleep quality in different sleep stages and throughout the night, will become available after validation. A comprehensive User manual provides detailed instructions for the use of the Viewer. Technical support is also available via phone/Internet to assist with installation and trouble-shooting.

Michele Sleep Scoring can be used with files generated by any data acquisition system provided the files can be exported in the EDF format.

System Requirements

1.1.1. The Viewer has been verified to work with, Windows® 7 and higher.

1.1.2. The Viewer requires a minimum of 4GB of RAM for guaranteed operation.

1.1.3. Processor requirement is 1.6 GHz or faster.

1.1.4. Monitor requirement 19 inch or greater, 1280 X 1024 resolution or greater.

1.1.5. Keyboard and mouse for control, a printer if hard copies are required.

1.1.6. 2 GB of free hard drive space.

1.1.7. Patient files must be in the .EDF format.

1.1.8. A new directory must be created for each patient file.

Data Collection Requirements

1. *Required Signals*: At a minimum, there should be two central EEG electrodes, one chin EMG and two eye electrodes for sleep staging and arousal detection. For scoring respiratory events, a minimum of nasal pressure (flow), Thermistor, Chest and Abdomen bands and SpO2 signals are required. End-tidal CO2 and audio are optional but are recommended for greater accuracy. Leg EMG signals are required if PLM scoring is to be performed.

2. *Electrode positioning and Impedance*: Electrode placement should follow the Standard Placement procedures (as per Rechtschaffen and Kales). Electrode impedance should be <5KΩ.

3. *Sampling rate*: The sampling rate should be a minimum of 120 Hz for AC channels (EEG, EOG, EKG, EMG and Raw, Unprocessed Audio), a minimum of 20 Hz for Respiratory signals (Thermister, Flow, Respiratory bands, Airway pressure, Airway CO_2), a minimum of 1 Hz for SpO_2 and a minimum of 0.03 Hz for body position, if position is recorded directly in the digital file.

4. *Signal filtering of AC channels* (EEG, EOG, EKG, EMG and Audio): Filtering must be performed either in hardware (at the amplifiers) or in the software after recording. If hardware filters are used, the file should be exported into the .EDF format without any additional filtering. If hardware filters are not used, it is recommended the file be exported without additional filters.

If the signals are to be filtered by hardware or during export, each of the 5 primary AC channels must be filtered as per AASM guidelines shown below. In addition, a 60 Hz notch filter must be used where available.

	Low Frequency Filter	High Frequency Filter
EEG	0.3 Hz	30-35 Hz (preferably 30)
EOG	0.3 Hz	30-35 Hz (preferably 30)
EMG	10 Hz	100 Hz
ECG	0.3 Hz	70 Hz
Audio	10 Hz	100 Hz

5. *Signal referencing*: As with filtering, referencing of AC signals must be performed either in the hardware (at the amplifiers) or in the software after recording.

If hardware referencing is used, the file should be exported into the .EDF format without any additional referencing. If hardware referencing is not used, it is recommended that the file is exported without referencing. Referencing can be done through the interface provided with the Viewer.

6. *Calibration*: All AC signals must be properly calibrated using a 50μV calibration switch.

Scoring Rules Used

The Michele Sleep Scoring System follows the most recent guidelines of the American Academy of Sleep Medicine (The AASM Manual for the Scoring of Sleep and Associated Events, 2007). Specifically:

1. *Identification of sleep stages*: This is done according to the Rechtschaffen & Kales rules, as modified by the new 2007 AASM guidelines, from analysis of the frequency profile of the EEG, presence of EEG spindles, k complexes and delta waves, level of chin EMG, and eye movements. Each 30-second epoch of data will be designated as being in stage W (awake), REM sleep, N1 Non-REM sleep, N2 Non-REM sleep, or N3 Non-REM sleep.

2. *Detection of arousals*: Each 30-second epoch is searched for the presence of one or more arousals following the guidelines of the American Sleep Disorders Association (Sleep 15: 174–184, 1992). Whenever an arousal is found, it will be classified as respiratory-related, leg movement-related or spontaneous, based on the temporal relation of the arousal to preceding events.

3. *Detection of respiratory events*:
Each 30-second epoch is searched for the presence of one or more respiratory events characterized by reduction (hypopnea) or complete cessation (apnea) of airflow in and out of the patient or Respiratory Effort Related Arousal (RERA). Whenever an apnea is detected, it will be characterized as central, obstructive or mixed. The software uses the guidelines

established by the AASM (The AASM Manual for the Scoring of Sleep and Associated Events published by the AASM in 2012). The user will have the choice of identifying hypopneas according to the Recommended or Alternate criteria of the 2012 AASM guidelines. Recommended criteria is defined as a reduction in airflow of at least 30% accompanied by an O2 desaturation of at least 3% or an associated sleep arousal. Alternate Criteria is defined as a reduction in airflow of at least 30% accompanied by an O2 desaturation of at least 4%. The software continuously evaluates the signal quality of the various respiratory channels (nasal cannula flow (or CPAP flow), thermistor, and chest and abdomen bands) and uses only signals that meet minimum quality criteria. If more than one channel is of acceptable quality, the amplitude criteria are assessed from nasal flow (or CPAP flow if patient is on CPAP), if valid. If not, the thermistor is used and if this is not valid also, the chest and abdomen bands signals are used to assess amplitude changes.

Hypopneas identified by either criterion are further classified into obstructive or undefined. An
obstructive hypopnea is one where there is associated snoring and/or chest wall paradox and/or flow limitation. An undefined hypopnea is one that lacks these features. It should be emphasized that lack of these features does not rule out an obstructive basis for the hypopnea. Hence, such hypopneas are called undefined, meaning they can be obstructive or central.

4. *Detection of leg movements*: Each 30-second epoch is searched for the presence of one or more leg movements. Whenever a leg movement is found, it will be characterized as associated or not with arousal and whether the leg movements follow a periodic pattern (periodic leg movements (PLMs)) following the guidelines established by American Academy of Sleep Medicine (The AASM Manual for the Scoring of Sleep and Associated Events published by the AASM in 2007).

Reported Variables

The report has two components, numerical and graphical. The Viewer contains a report editor utility that allows custom reports to be built that mirror closely those normally generated by clinics; i.e., the reports will look the same as clinics are used to but will be automatically populated with patient data by Michele. The Viewer also comes with default reports that contain all of the values and graphs normally associated with a PSG study. Even with these default reports, customization is possible, as clinics utilizing Michele may opt to exclude specific fields and graphs from appearing on the report; the default is for all fields and graphs to be displayed in the default reports. In the case of a split study, there will be two numerical reports – one for the pre-CPAP period and one for the On-CPAP period.

Validation Results

There were one completed internal and one completed external validation studies. The external study was conducted by well-established academic investigators who independently designed the study, selected the files and performed the comparisons

between the automatic and manually scored results. Younes Medical Technologies' role was to simply score the files ***on site*** and deliver the results immediately to the investigators. Other studies are underway to determine what needs to be edited, the time required for editing and the impact of editing on the clinically relevant results.

***Internal Validation*:**

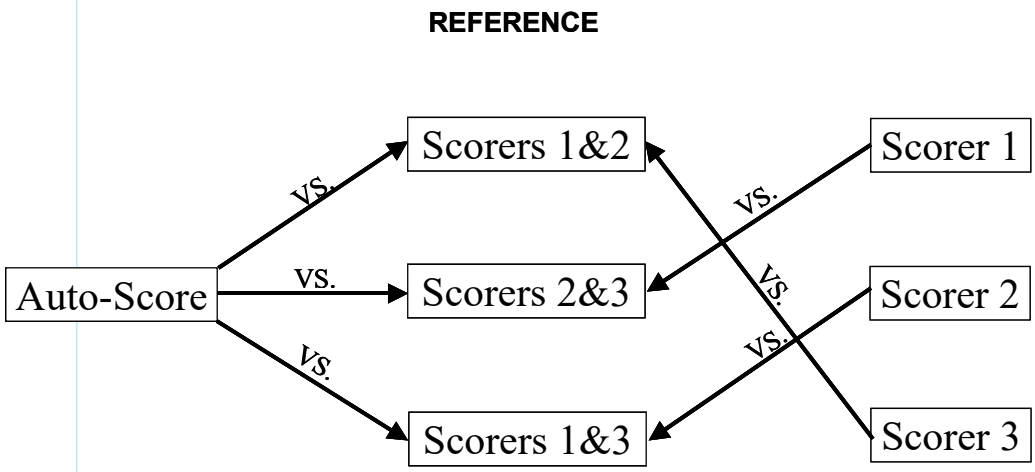
The software performance was measured by: A) Determining epoch-by-epoch agreement between Michele's scoring and the scoring of three technologists with respect to the four scoring functions (Sleep staging, Arousals, Periodic Leg Movements (PLMs) and Respiratory Events) (Objective 1). B) Determining the agreement between Michele's results of Clinically Relevant Data, such as Total Sleep Time, Time in Different Stages, Apnea and Hypopnea Index (AHI…etc) and the results of the three technologists (Objective 2).

1. Files: Software performance was assessed using 30 full night studies recorded in the sleep laboratory of a tertiary care facility (Foothills Hospital, Calgary, Canada). The files were selected at random and included 19 patients with sleep apnea. Fifteen of these patients had moderate to severe sleep apnea (AHI 73±38 hr^{-1}) and underwent split studies with one part (pre-CPAP) where sleep was severely fragmented and a second part (on CPAP) with fairly normal sleep and breathing. The group also included 9 patients with PLMs (8 to183 hr^{-1}; average 38±55 hr^{-1}), two patients with severe sleep fragmentation for no apparent cause (non-organic insomnia) and seven patients with normal sleep. Overall, the quality of sleep varied considerably among the 30 patients with Total Sleep Time ranging from 2.6 to 7.8 hours (4.2±1.1 hours), sleep efficiency ranging from 37 to 99% (61±18%) and arousal index ranging from 9 to 97 hr^{-1} (17±4 hr^{-1}). A total of 24967 thirty-second epochs were scored.

2. Technologists: Each of the three scorers is Board certified and has had at least 15 years of hands-on experience in scoring polysomnograms.

3. Analytical Methods and Results:

3.1. *Objective 1 testing* (epoch-by-epoch agreement): Comparisons between two qualified technologists usually reveal differences in the scoring in 15-25% of the epochs. Accordingly, differences in scoring between the automatic scoring and that of individual human scorers will inevitably show differences and it is not possible through such comparisons to determine who is right and who is wrong. Accordingly, for this validation study we used a design that permits the direct comparison between the auto-scoring and single scorers. This design is shown in figure 1. The scoring of scorers 1 and 2 were compared. Epochs in which both scorers agreed on the score were noted and the score in these epochs was used as a reference for comparison with scorer 3 and the automatic scoring. The % agreement between the auto-score and the reference is compared with the % agreement between the third scorer and the reference. In this fashion the auto-score and the third scorer are both compared to the same reference. The same is done using as reference the scores of another pair (2 & 3) and testing the auto-score vs. scorer #1 when



both are compared to the same reference, and so on for the third pair. In this fashion, the auto-score was compared directly with each of the three experienced scorers.

Figure 1: Design of the Internal validation study

Table 1 shows the epoch-by-epoch agreement between the auto-score and the consensus of scorers 1 and 2 (left panel) and between scorer 3 and the consensus of scorers 1 and 2 (right panel). It can be seen that the overall % agreement for epochs in which there was agreement between scorers 1 and 2 was the same for the auto-score (88.2%) and for scorer 3 (also 88.2%). Scorers 1 and 2 did not agree on the score in 16.7% of the epochs. In these epochs at least one agreed with the auto-score in 88.9% of these ambiguous epochs (N+/Total N).

	S1 + S2						
	Epochs with Agreement					No Agreement	
Auto	**Awake**	**N1**	**N2**	**N3**	**Rem**	**N-**	**N+**
Awake	**5119**	121	183	57	55	90	853
N1	148	**750**	471	5	134	116	887
N2	117	448	**7445**	78	163	174	1163
N3	9	0	387	**2479**	0	20	666
Rem	26	26	37	10	**2760**	70	183
Total	5419	1345	8523	2629	3112	470	3752
PPA	94.5	55.8	87.4	94.3	88.7		
NPA	97.3	96.1	93.6	97.8	99.4		
% agreement	88.2						
kappa (%)	83.9						
N / total (%)	16.7						
N+ / Total N (%)	88.9						

	S1 + S2						
	Epochs with agreement					No Agreement	
S3	**Awake**	**N1**	**N2**	**N3**	**Rem**	**N-**	**N+**
Awake	**5336**	190	83	10	84	99	1144
N1	58	**990**	576	11	266	143	1066
N2	22	145	**7679**	776	39	27	1323
N3	0	0	122	**1807**	0	1	233
Rem	3	20	63	0	**2723**	13	173
	5419	1345	8523	2604	3112	283	3939
	98.5	73.6	90.1	69.4	87.5		
	97.6	95.4	92.1	99.3	99.5		
	88.2						
	83.8						
	16.7						
	93.3						

Table 1: Epoch-by-epoch agreement between Auto and S1/S2 (left) and between S3 and S1/S2

Table 2 shows the results for Auto vs. S1, both compared to S2/S3. Here, Auto had slightly lower agreement than S1 (85.4% vs. 88.3%).

	S2 + S3						
	Epochs with agreement					No Agreement	
Auto	**Awake**	**N1**	**N2**	**N3**	**Rem**	**N-**	**N+**
Awake	**5746**	93	152	43	34	92	318
N1	389	**778**	373	1	97	74	799
N2	241	533	**7150**	56	176	201	1231
N3	20	6	682	**1779**	1	35	1038
Rem	63	44	41	15	**2489**	53	407
Total	6459	1454	8398	1894	2797	455	3793
PPA	89.0	53.5	85.1	93.9	89.0		
NPA	97.8	95.6	92.0	96.3	99.1		
% agreement	85.4						
kappa (%)	79.8						
N / total (%)	16.8						
N+ / Total N (%)	89.3						

	S2 + S3						
	Epochs with agreement					No Agreement	
S1	**Awake**	**N1**	**N2**	**N3**	**Rem**	**N-**	**N+**
Awake	**5336**	1	3	0	0	2	104
N1	898	**990**	206	0	14	61	957
N2	187	431	**7679**	87	60	165	1448
N3	14	5	489	**1807**	0	17	968
Rem	24	27	21	0	**2723**	38	488
	6459	1454	8398	1894	2797	283	3965
	82.6	68.1	91.4	95.4	97.4		
	100.0	94.3	93.9	97.3	99.6		
	88.3						
	83.8						
	16.8						
	93.3						

Table 2: Epoch-by-epoch agreement between Auto and S2/S3 (left) and between S1 and S2/S3

Table 3 shows result of auto vs. S2, when both are compared to S1/S3. Here, Auto's performance was also slightly lower than S2.

	S1 + S3						
	Epochs with agreement					No Agreement	
Auto	**Awake**	**N1**	**N2**	**N3**	**Rem**	**N-**	**N+**
Awake	**5076**	181	167	44	38	127	845
N1	136	**836**	452	1	97	107	882
N2	112	527	**7165**	54	145	167	1418
N3	9	1	474	**1844**	0	29	1204
Rem	24	47	25	10	**2542**	68	396
Total	5357	1592	8283	1953	2822	498	4745
PPA	94.8	52.5	86.5	94.4	90.1		
NPA	97.1	96.3	92.9	97.3	99.4		
% agreement	87.3						
kappa (%)	82.5						
N / total (%)	20.8						
N+ / Total N (%)	90.5						

	S1 + S3						
	Epochs with agreement					No Agreement	
S2	**Awake**	**N1**	**N2**	**N3**	**Rem**	**N-**	**N+**
Awake	**5336**	363	69	1	32	120	1206
N1	16	**990**	308	0	46	68	819
N2	3	222	**7679**	145	20	56	1563
N3	0	0	201	**1807**	1	4	909
Rem	2	17	26	0	**2723**	35	463
	5357	1592	8283	1953	2822	283	4960
	99.6	62.2	92.7	92.5	96.5		
	96.8	98.0	96.7	98.9	99.7		
	92.6						
	89.8						
	20.8						
	94.6						

Table 3: Epoch-by-epoch agreement between Auto and S1/S3 (left) and between S2 and S1/S3

Table 4 is the average of tables 1 to 3. It shows that the agreement between auto-scoring and the consensus of two expert scorers for sleep staging is 87%, slightly less than the average of the three individual scorers vs. the same reference values (89.7%). It also shows that in 89.6% of epochs where two scorers do not agree, at least one of the two agrees with the auto-score (N+ / Total N).

	Average all 3 Auto vs. 3 pairs							Average all 3 techs vs. 3 pairs						
	Awake	**N1**	**N2**	**N3**	**Rem**	**N-**	**N+**	**Awake**	**N1**	**N2**	**N3**	**Rem**	**N-**	**N+**
	5314	132	167	48	42	103	672	5336	185	52	4	39	74	818
	224	788	432	2	109	99	856	324	990	363	6	109	91	947
	157	503	7253	63	161	181	1271	71	266	7679	340	40	83	1445
	13	2	514	2034	0	28	969	5	2	271	1809	0	7	703
	38	39	34	12	2597	64	329	10	21	37	0	2723	29	375
Total	5745	1464	8401	2159	2910	474	4097	5745	1464	8401	2159	2910	283	4288
PPA	92.5	53.9	86.3	94.2	89.3			93.6	68.0	91.4	85.8	93.8		
NPA	97.4	96.0	92.8	97.2	99.3			98.1	95.9	94.2	98.5	99.6		
% agreement	87.0							89.7						
kappa (%)	82.1							85.8						
N / total (%)	18.1							18.1						
N+ / Total N (%)	89.6							93.7						

Table 4: Epoch-by-epoch agreement between Auto and three pairs of scorers (left) and between each of the 3 scorers vs. the same three pairs.

Tables 5, 6 and 7 show results for arousal scoring, PLM scoring and respiratory event scoring. The % agreement was quite high in all cases and quite comparable to the results of individual scorers vs. the same reference.

AROUSALS

	Average Auto vs. 3 pairs				Average human scorer vs. 3 pairs			
	Yes	**No**	**N-**	**N+**	**Yes**	**No**	**N-**	**N+**
Yes	932	637	43	783	1101	610	76	945
No	527	14006	70	1374	394	14095	38	1221
Total	1458	14646	113	2164	1496	14705	114	2166
PPA	63.3				73.6			
NPA	95.8				95.9			
% agreement	92.9				93.8			
kappa (%)	57.6				66.0			
Total N	2276				2280			
N / total (%)	12.0				13.3			
N+ / Total N (%)	95.1				95.1			

Table 5: Epoch-by-epoch agreement for arousals between auto-score and three pairs of scorers (left) and between each of the 3 scorers vs. the same three pairs.

PLMs

Average Auto vs. 3 pairs				No Agreement	
Auto	1 PLM	2 PLMs	None	N-	N+
1 PLM	506	23	280	12	317
2 PLMs	21	252	28	15	70
None	107	33	16251	12	444
Total	635	308	16559	39	832
PPA (1&2 PLMs)	85.1				
NPA (1&2PLMs)	98.1				
% agreement	97.2				
kappa (%)	74.8				
Total N	871				
N / total (%)	5.0				
N+ / Total N (%)	95.5				

Average Tech. vs. 3 pairs				No Agreement	
Tech	1 PLM	2 PLMs	None	N-	N+
1 PLM	508	22	226	16	276
2 PLMs	18	261	18	16	83
None	120	19	16393	16	487
	646	303	16637	49	847
	85.3				
	98.5				
	97.6				
	77.9				
	871				
	4.9				
	97.2				

Table 6: Epoch-by-epoch agreement for PLMs between auto-score and three pairs of scorers (left) and between each of the 3 scorers vs. the same three pairs.

RESPIRATORY EVENTS

Average Auto vs 3 pairs						No Agreement	
Auto	H	O	M	C	0	N+	N-
H	895	43	1	14	360	478	30
O	75	114	2	3	18	122	16
M	11	13	139	10	8	42	29
C	8	7	26	85	17	49	38
0	111	9	0	14	15380	269	57
Total	1100	185	167	126	15783	960	170
PPA	81.3	61.3	83.0	67.5	97.4		
NPA	97.4	99.4	99.8	99.7	91.5		
%agreement	95.7						
kappa (%)	76.4						
Total N	1130						
N / total (%)	6.1						
N+/Total N (%)	84.9						

Average Tech. vs 3 pairs						No Agreement	
Tech	H	O	M	C	0	N+	N-
H	917	53	1	4	203	395	24
O	75	111	11	1	14	146	18
M	2	9	149	4	8	46	11
C	2	1	5	96	18	56	13
0	121	11	7	20	15522	383	56
	1117	185	173	124	15765	1026	121
	82.1	59.8	86.1	76.9	98.5		
	98.4	99.4	99.9	99.8	90.1		
	96.7						
	81.4						
	1147						
	6.2						
	89.5						

Table 7: Epoch-by-epoch agreement for respiratory events between auto-score and three pairs of scorers (left) and between each of the 3 scorers vs. the same three pairs.

3.2. *Objective 2 testing*:

In this section we discuss agreement between automatic and manual scoring for summary variables that appear in the clinical report used by physicians to assess sleep disorders. Table 8 shows the results for 14 variables. These were selected because they are the most commonly used variables in the clinical assessment.

TABLE 8 Variable	Average S1-S3 SD	Michele SD	Michele - ave. SD	Intra-Class Correlation Michele vs. Ave.	S1 vs. Ave.	S2 vs. Ave.	S3 vs. Ave.
Total sleep time (min)	312 74	312 72	0 13	0.983	0.954	0.978	0.992
Sleep efficiency (%)	74.6 17.0	74.7 16.6	0.1 2.9	0.985	0.957	0.98	0.994
Sleep-onset latency (min)	24 27	24 29	0 9	0.950	0.991	0.997	0.995
REM-onset latency (min)	126 71	126 73	0 28	0.923	0.988	0.966	0.992
Stage wake (min)	108 76	108 74	-1 12	0.986	0.958	0.98	0.994
Stage 1 (min)	47 30	42 28	-5 14	0.876	0.912	0.864	0.91
Stage 2 (min)	165 54	159 50	-5 20	0.922	0.983	0.964	0.972
Stage 1+ 2 (min)	212 57	202 62	-10 21	0.923	0.951	0.935	0.95
Stage delta (min)	47 38	60 46	13 17	0.869	0.940	0.847	0.948
Stage REM (min)	53 26	51 26	-2 8	0.951	0.988	0.977	0.984
Arousal Index (hr^{-1})	33 23	25 11	-9 15	0.566	0.937	0.956	0.932
PLM Index (hr^{-1})	12 29	13 31	1 9	0.958	0.978	0.855	0.867
AHI A (hr^{-1})	30 41	32 40	2 7	0.982	0.992	0.974	0.988
AHI B (hr^{-1})	31 42	27 36	-4 8	0.971	0.99	0.967	0.986
Average				0.918	0.966	0.946	0.965

The first data column of Table 8 is the average score of the three technologists for each of the 14 variables of interest. The averaging was done on a file-by-file basis. The values and corresponding standard deviations (SD) given in column 1 are the average and SD of the 30 averages. The second column contains the average and SD of the values obtained from Automatic analysis with Michele Sleep Scoring System. The third column lists the average and SD of the thirty differences between Michele and the corresponding average of the three technologists (Bland and Altman analysis). The last four columns contain the intra-class correlation coefficients for comparisons between each of the four scorers (Auto and three scorers) and the average score of the three technologists.

The results show good agreement in general between Michele scores and the average of three technologists. With the exception of the arousal index where concordance (ICC) between the Auto-score and the average was only modest (ICC = 0.566), concordance was excellent and mostly within the range observed in comparisons between individual technologists and the average of the three technologists. Average ICC for Michele vs. Average was 0.918 (bottom row, table 3), only marginally below S1 (p=0.03 by ANOVA for repeated measures) and not significantly different from S2 or S3. The low correlation for arousal scoring was related to a coding error that affected one of the files and has since been corrected.

***External Validation*:**

1) **The Alliance Study**: The Academic Alliance for Sleep Research (AASR) was undertaking a study to investigate the scoring variability within and across five major academic sites. The sites involved were the University of Pennsylvania, Harvard University, Stanford University, University of Wisconsin and St. Lukes Medical Center, MO. The Alliance invited Magdy Younes, MD, president of Younes Medical Technologies, to participate in this research, while the Alliance maintained control of study design, file selection, data analysis and interpretation. Seventy PSG files of good quality were selected by the University of Pennsylvania Sleep Center (UPSC) and distributed to the five US academic sleep centers to be scored by two technologists at each center. All scorers had at least 4 years of PSG scoring experience. Sleep staging and arousals were scored according to the 2007 American Academy of Sleep Medicine (AASM) guidelines. Respiratory events were scored by 3 different criteria: AASM recommended, AASM alternate, and AASM clinical research (Chicago) criteria. All manually scored data were transferred to UPSC via Internet. Automatic scoring was performed at UPSC by a Younes Medical Technologies technician who analyzed the 70 PSG files using a laptop containing the software. No manual editing of the automatic scoring was performed. The results of the automatic scoring were immediately delivered to personnel at UPSC and a UPSC bio-statistician performed the analysis. For each PSG variable examined (see Table), the values obtained with the automatic scoring software were compared to the average score of the 10 technologists. Intra-class correlation (ICC) was performed on the 70 pairs of data and was compared to the Across-Sites ICC. The latter was calculated from the average values of the two technologists at each site.

Table 9 shows the results. The table shows that the agreement between the automatic score and the average of the 10 manual scores was comparable to the agreement across sites. Variables that showed high ICC across sites also showed high ICC for Auto vs. Average. The

ICCs of automatic versus manual scoring and of manual scoring across 5 sites

Variable	Auto vs. Average	Across Sites	Diff.	Variable	Auto vs. Average	Across Sites	Diff.
Total Sleep Time (min.)	0.865	0.870	-0.005	AHI "Recommended"	0.960	0.980	-0.020
Sleep Efficiency (%)	0.738	0.767	-0.029	AHI "Alternate"	0.910	0.950	-0.040
Stage N1 (minutes)	0.559	0.441	0.118	AHI "Chicago criteria"	0.750	0.800	-0.050
Stage N2 (minutes)	0.844	0.614	0.230	Rem AHI Criteria A	0.950	0.730	0.220
Stage N3 (minutes)	0.472	0.402	0.070	Rem AHI Criteria B	0.918	0.678	0.240
Stage Rem (minutes)	0.637	0.685	-0.048	Rem AHI Chicago	0.725	0.590	0.135
# Arousals, NREM	0.834	0.575	0.259	# Central Apneas	0.625	0.683	-0.058
# Arousals, REM	0.388	0.522	-0.134	# Obstructive Apneas	0.995	0.861	0.134

ICC for Auto vs. Average was higher than the ICC across sites in 8 of the
Table 9: Results of the Alliance Study

16 variables, sometimes by a large amount (e.g. Stage N2, non-Rem arousals and Rem AHI, A and B). When the ICC for Auto vs. Average was lower, the differences were very small. It was concluded that this automatic PSG yielded results similar to those obtained by experienced technologists using computer-assisted manual scoring.

The results of this study were presented at the American Thoracic Society (ATS) meeting in San Francisco, May 2012 and published in the Abstract Edition of volume 185 of the *American J of Respir Crit Care Med* (Abstract A6428). A manuscript describing these findings has been published in the April 2013 edition of the journal *SLEEP* (Malhotra A, Younes M, Kuna ST, Benca R, Kushida CA, Walsh J, Hanlon A, Staley B, Pack AI, Pien GW. Performance of an Automated Polysomnography Scoring System Versus Computer-Assisted Manual Scoring: http://www.journalsleep.org/ViewAbstract.aspx?pid=28888).

2) **Other Studies**: Studies are currently underway in multiple centers to determine the extent of editing required and its impact on clinically relevant variables. The protocol consists of comparing scoring results from the automatic system before and after editing by highly qualified technologists. Editing is done in two ways: A) Epoch by epoch, correcting sleep stage and events whenever the technologist disagrees with the automatic score and with no time limit. B) Expedited editing where the technologists checks only for certain scoring outcomes and does a very quick scan of the rest of the file. Preliminary data so far indicate that detailed editing consumes on average about 30 minutes but does not result in any clinically significant changes in the report except for sleep and REM onset latencies and the very rare case where one or more periods of REM sleep are missed because of sub-optimal chin EMG signal. The expedited editing focuses on these potential errors and can be done in 5 to 10 minutes.