Clinical Validation of EnsoSleep AI Scoring in Adult and Pediatric Patients Compared to a Prospective, Double-Blind Scoring Panel

Research Contributors:

Chris R. Fernandez, MS1, Sam Rusk, BS1, Nick Glattard, MS1, Yoav N. Nygate, MS1, Fred Turkington, BS1, Justin Mortara, PhD1, Nathaniel F. Watson, MD, MS2

Introduction

Artificial intelligence (AI) and more specifically, machine learning (ML) and deep learning (DL), have shown great promise to help unlock clinical insights hidden in healthcare data and medical waveforms. Since before our initial FDA clearance in 2017, our team has been working to bring the promise of AI technology to life for clinicians and patients around the world.

An important step on that journey is to validate that an AI/ML engine can improve its performance over time as it gains experience from new clinical data. In this case, clinicians who use the application provide feedback through editing and scoring review that can be used for testing, system monitoring, and ultimately to study and to improve the AI/ML engine performance and generalizability.

Since EnsoData's AI/ML engine for sleep staging and event detection (EnsoSleep) was FDA cleared in 2017, more than 410,000 sleep studies have been analyzed by the Waveform AI engine to aid clinicians in diagnosing sleep disorders and sleep apnea. In 2020, we set out to validate that EnsoSleep had improved our Waveform AI, by conducting a study with roughly three times the clinical subject sample size than our prior study and other studies of its kind.

The dataset used in the clinical validation is large, diverse, and representative, with a wide variety of demographic differences, including age, gender, BMI, medical conditions, medications, and other characteristics. We set new benchmarks for AI scoring including: performance for Sleep Staging, OSA, CSA, Hypopneas, Arousals, Limb Movements, Cheyne-Stokes Respiration, and Periodic Breathing Episodes, plus normative, mild, moderate, and severe OSA categories for overall AHI and AHI during REM periods.

EnsoSleep, which is compatible with most PSG and HSAT hardware/software platforms and can be implemented with limited or no impact on existing clinician workflows, has the potential to free up clinical staff to perform other duties and improve the end-to-end sleep care experience.

Methods

A semi-prospective, randomized, cross-sectional study design was utilized to construct a representative sample of patient PSG data collected directly from the intended use population by qualified intended users in the clinical laboratory setting. The following sections describe the study protocol.

Laboratory Selection

Controls were placed on laboratory selection. Laboratories were required to:

- maintain current AASM Sleep Testing Facility Accreditation
- have multiple regional sleep center facilities (e.g. be a multi-site laboratory)
- maintain an archived collection of clinical diagnostic PSG records that includes a complete spectrum of subject disease states, relevant medical conditions, and demographics, as detailed in Table 1, (beginning on this page, and continuing on the following page).

Five (5) clinical testing multi-site laboratories were evaluated, and an AASM Accredited Sleep Testing Facility with two (2) regional sleep testing centers was selected for this study as meeting all laboratory quality, external validity, and subject spectrum controls.

Sample Demographics	EnsoSleep Adult Sample (n=100)	EnsoSleep Pediatric Sample (n=100)
Age (years)	**60.21 ± 15**	**11.99 ± 4**
5-8	0	25
9-12	0	28
13-15	0	18
16-17	0	36
18-29	4	0
30-39	10	0
40-49	10	0
50-59	18	0
60-69	26	0
70-79	22	0
80-99	10	0
Gender		
Female	46	40
Male	54	60

Adult Sample Demographics	EnsoSleep Adult Sample (n=100)	EnsoSleep Pediatric Sample (n=100)
BMI		
Underweight	1	34
Normal	3	6
Overweight	50	40
Obese	33	15
Morbidly Obese	13	5
AHIPSG (events/hour)	**22.36 ± 18.80** (range 0 - 109.22)	**8.89 ± 10.67** (range 0 - 60.25)
Disease Severity		
Normative	17	51
Mild	23	34
Moderate	32	12
Severe	28	3
Relevant Medical Condition Groups		
Sleep Disorders	83	49
Psychiatric Disorders	7	5
Neurologic Disorders	6	3
Neurodevelopmental Disorders	0	8
Cardiac Disorders	61	9
Pulmonary Disorders	24	29
Metabolic and Other Disorders	13	1
Relevant Medication Groups		
Benzodiazepines	6	1
Antidepressants	14	18
Stimulants	1	7
Opiates	6	0
Sleep Aids	2	0

Table 1. Demographic data for 2021 EnsoSleep clinical validation samples.

The archived collection was completely separate from any and all PSG data used previously for EnsoSleep device development, software verification testing, software validation testing, or production usage, and was obtained independently for validation and specifically for clinical validation performance testing.

Study Sample Selection

Once all laboratory selection control requirements were met, an archived collection of N=1,984 PSG records acquired from 2018 through 2020 was gathered. We used the AASM Digital Task Force evidence grading framework to determine an appropriate study sample size from the archived collection. A sample size of N=100 subjects was determined to meet the highest grade, level one (1) performance criteria.

A randomized sampling procedure was applied to the archived collection to determine an Adult and Pediatric Sample that showed no statistically significant difference in disease state distributional characteristics according to a two-sided t-statistic based 95% confidence interval relative to the archived collection study adult and pediatric populations.

Comparative Reference

We designated 2/3 Majority Scoring, defined as manual scoring by a panel consensus among three (3) double-blinded, registered sleep technologists following the rules and recommendations outlined in the AASM Manual, as a reference standard to compare performance against EnsoSleep AI scoring.

To control for the event-dependent variability in scoring agreement between raters, a cross-sectional study design was utilized to introduce several controls:

- **Acquisition-blind rater controls:** all scoring technologists were blinded from acquisition technologist notes and interaction
- **Scoring-blind rater controls:** each rater was blinded from any and all scoring technologist notes and interaction with respect to an individual subject
- **Rater quality controls:** all independent raters utilized in either data acquisition or manual scoring must maintain current professional certification of RST, RPSGT, CPSGT, CRT-SDS or

Six (6) clinical testing laboratories were recruited for independent manual scoring, and a clinical test setting with N=9 total registered scoring technologists ranging from five (5) to twenty (20) years of clinical sleep experience, meeting the above controls.

Manual scoring for Sleep Stages, Obstructive Apneas, Central Apneas, Hypopneas, Respiratory Effort Related Arousals, Arousals, Limb Movements, Cheyne-Stokes Respiration Episodes,

and Periodic Breathing Episodes were obtained from three (3) independent registered sleep technologists (RPSGT) for the N=100 adult and pediatric subjects selected as the final study samples.

To construct the 2/3 Majority Scoring, a manual scoring reference was derived for each patient by computing the 30 second epochs of which at least two (2) raters agreed on the presence of a given event type (for example, a given epoch should be scored as containing REM sleep and a hypopnea event).

Statistical Analyses

To assess sleep staging performance, an epoch-by-epoch 2/3 Majority Scoring comparison to EnsoSleep was conducted for total and individual Sleep Stages (W/N1/N2/N3/R) and all sleep events (Obstructive Apneas, Central Apneas, Hypopneas, Respiratory Effort Related Arousals, Arousals, Limb Movements, Cheyne-Stokes Respiration Episodes, and Periodic Breathing Episodes).

Positive Agreement, Negative, Agreement, Overall Agreement (PA/NA/OA), and inter-rater Cohen's Kappa (K) coefficients, with two-sided, positive and negative, 95% confidence intervals (CI) using the Bootstrap Percentile method with R=1,000 resampling were calculated.

- Overall percent agreement = 100% x (TP+TN) / (TP+FP+TN+FN)
- Positive percent agreement = 100% x TP / (TP+FN)
- Negative percent agreement = 100% x TN / (FP+TN)

Where,

TP = number of true positive events
FP = number of false positive events
TN = number of true negative events
FN = number of false negative events

Reference Standard

	Condition Present	Condition Absent
	+	-
	TP	FP
	FN	TN
	TP+FN	FP+TN

Graphic 1: Reference PA, NA, & OA: Source

In the case of OSA diagnostic agreement, overall and REM AHI were computed and analysis was conducted on two predefined diagnostic thresholds: AHI ≥ 5 and AHI ≥ 15, representing normative versus mild sleep apnea and mild versus moderate sleep apnea respectively, and

the same 2/3 Majority Scoring comparison to EnsoSleep was conducted to determine PA/NA/OA, and Cohen's Kappa (K) coefficient with two-sided, bootstrapped 95% CIs.

For diagnostic agreement, positive and negative diagnostic likelihood ratios are calculated as an additional statistic. [1] Positive likelihood ratios correspond to the clinical concept of "ruling-in disease,"and can be interpreted as the ratio of probability that a subject with sleep apnea has a positive AHI result to the probability that a subject without sleep apnea has a positive AHI result.

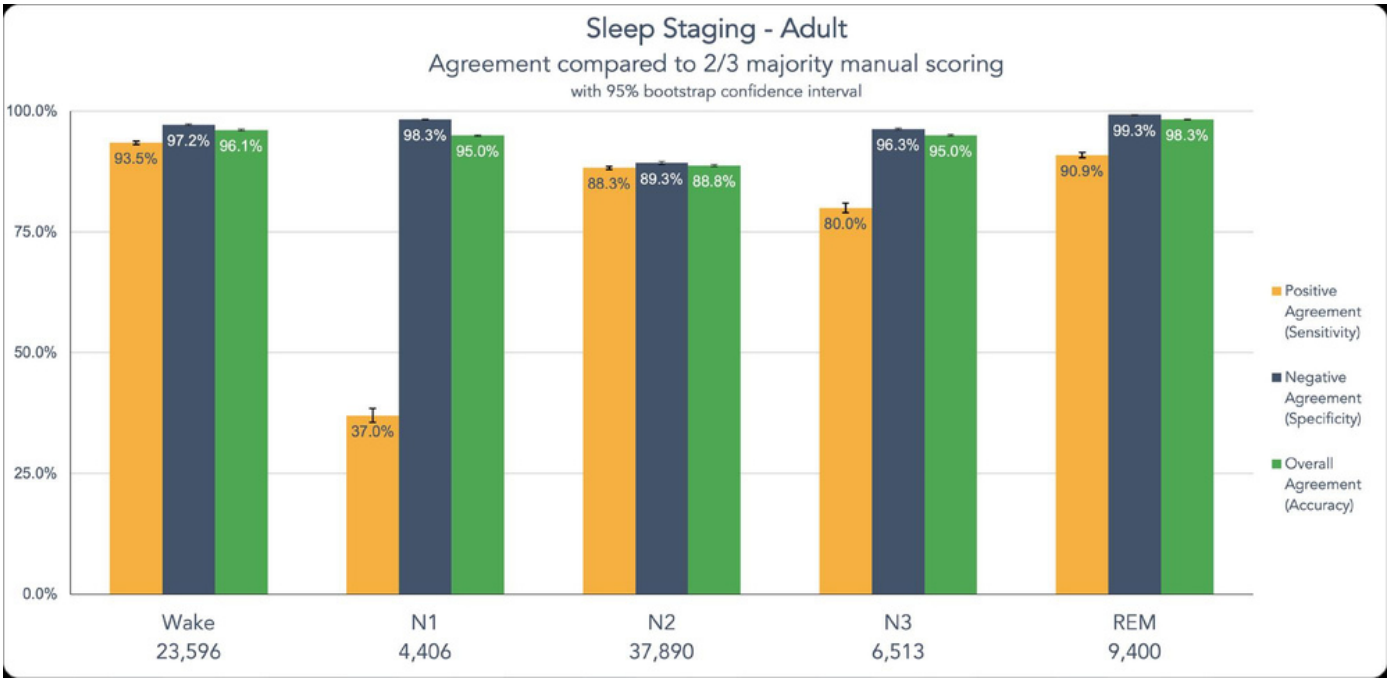
Study Endpoints

EnsoSleep device performance was evaluated using the defined cross-sectional experimental design, statistical methodology, and set of comprehensive experimental controls, across the following four (4) experimental endpoints:

- Endpoint 1: EnsoSleep is intended to assist clinicians with the assessment of sleep quality, therefore performance of device sleep scoring must be validated.
- Endpoint 2: EnsoSleep is intended to assist clinicians with the scoring sleep disordered breathing events used in diagnostic evaluation, therefore device performance for diagnosing sleep apnea must be validated.
- Endpoint 3: EnsoSleep is intended to analyze physiological signals and automatically score sleep study results, including detection of SDB events, Hypopnea events, Apnea events, including OSA events, CSA events, Arousal events, Limb Movement events, RERA events, CS events, and PB events, therefore device performance for detecting each event type must be validated.
- Endpoint 4: EnsoSleep is intended to analyze physiological signals and automatically score sleep study results, including detection of Respiratory Rate, sleep vs. wake stages in photoplethysmogram signal (PPG), and resulting Total Sleep Time (TST), therefore device performance for detecting each event type must be validated.

Results - Adult Sample

Study Endpoint 1 - Adult Sleep Staging



Graphic 2. The results of the epoch-by-epoch comparison to 2/3 Majority manual scorers in the Adult Sample (N=100) are presented.

Sleep Stage	2021 Positive Agreement	2017 Positive Agreement	2021 Kappa (95% CI)	2017 Kappa (95% CI)
W/N1/N2/N3/R	87%	78%	.825 (.823-.828)	.73 (.72-.73)
Wake	94%	86%	.891 (.888-.895)	.84 (.84-.85)
N1	37%	41%	.380 (.366-.394)	.30 (.29-.31)
N2	88%	77%	.752 (.748-.756)	.65 (.64-.66)
N3	80%	81%	.693 (.684-.702)	.61 (.60-.62)
REM	91%	79%	.907 (.902-.912)	.81 (.80-.82)

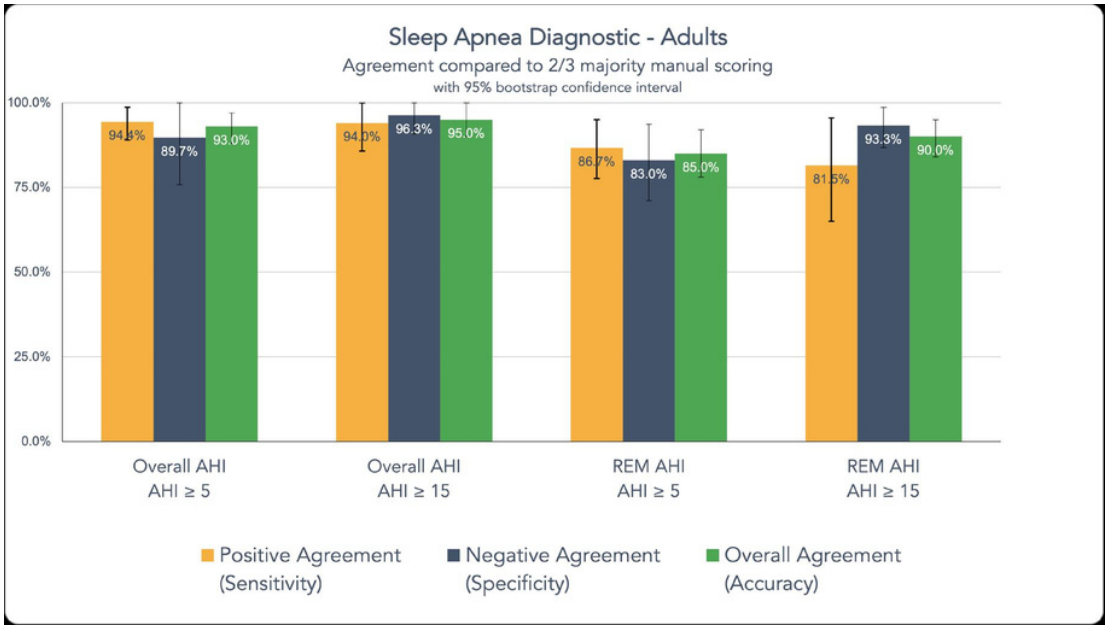
Table 2. Positive Agreement, or the % of studies where EnsoSleep and 2/3 Majority manual scorers agreed on sleep stage, and Cohen's Kappa (K) for EnsoSleep in 2017 and 2021 are presented.

For further reading on sleep medicine clinician inter-scorer reliability performance benchmarks in a large sample of epochs for many of the sleep stages and scoring events discussed here, we the work presented by Rosenberg & Van Hout. The AASM inter-scorer reliability program [2,3].

		EnsoSleep Staging				
2/3 Majority Expert Manual Scoring	Sleep Stage	Wake	N1	N2	N3	REM
	Wake	93.5%	2.9%	3.4%	0.0%	0.1%
	N1	18.2%	37.0%	43.8%	0.0%	0.9%
	N2	1.8%	1.4%	88.3%	7.3%	1.2%
	N3	0.2%	0.0%	19.8%	80.0%	0.0%
	REM	1.0%	0.9%	7.1%	0.0%	90.9%

Table 3. Orange cells indicated percent of epochs where EnsoSleep and 2/3 Majority manual staging were in agreement.

Study Endpoint 2 - Adult Sleep Apnea Diagnostic



Graphic 3. Adult Sleep Apnea Diagnostic Data Compared to 2/3 Majority Scoring.

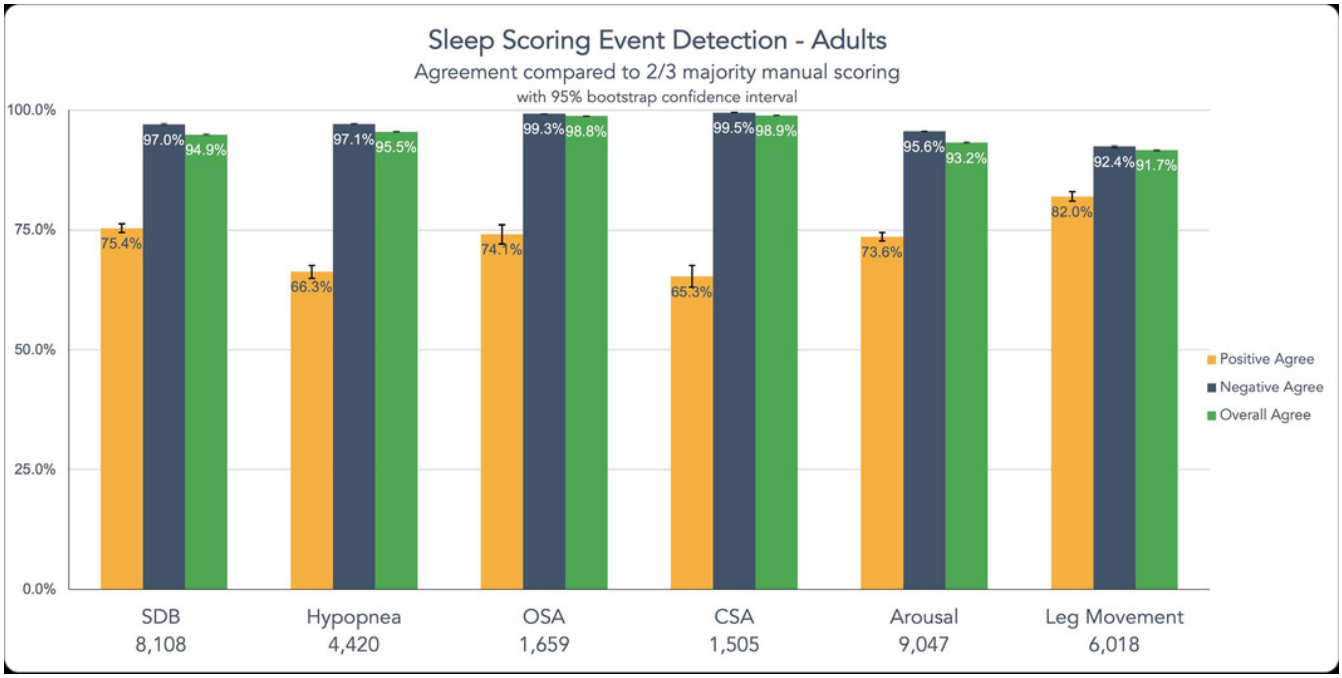
Likelihood Ratio	Overall AHI AHI > 5	Overall AHI AHI > 15	REM AHI AHI > 5	REM AHI AHI > 15
Likelihood Ratio (+)	9.146	25.458	5.069	12.052
95% bootstrap CI	3.879, ∞	10.154, ∞	2.692, 13.597	5.977, 55.250
Likelihood Ratio (-)	0.062	0.062	0.162	0.198
95% bootstrap CI	0.014, 0.127	0.000, 0.151	0.060, 0.278	0.049, 0.384

Table 4. The results of the comparison of OSA severity to 2/3 Majority scorers are presented.

Sleep Apnea Diagnostic	2021 Positive Agreement %	2017 Positive Agreement %
AHI ≥ 5	94%	91%
AHI ≥ 15	94%	95%
REM AHI ≥ 5	87%	83%
REM AHI ≥ 15	82%	79%

Table 5. Positive Agreement, or the % of studies where EnsoSleep and 2/3 Majority results in the same diagnostic category, and Cohen's Kappa (K) for EnsoSleep in 2017 & 2021 are presented.

Study Endpoint 3 - Adult Sleep Event Detection



Graphic 4. The results of sleep event detection for common sleep event types are presented. Positive Agreement was between 65.3% - 82% across these common event types.

Event Type (Adults)	2021 Positive Agreement %	2017 Positive Agreement %
Sleep Disordered Breathing (SBD)	75%	67%
Hypopnea	70%	60.3%
Apnea	73%	56%
OSA	74%	53%
CSA	65%	63.8%
Arousal	74%	66%
Limb Movement	82%	71%
RERA	35%	n/r
Cheyne-stokes respiration episode (CSE)	47%	n/r

Table 6. Positive Agreement, or the % of studies where EnsoSleep and 2/3 Majority agreed on event detection, and Cohen's Kappa (K) for EnsoSleep in 2017 and 2021. n/r = not recorded).

Study Endpoint 4 - Total Sleep Time and Respiratory Rate

The following is a table that compares the EnsoSleep scoring samples against a 2/3 majority.

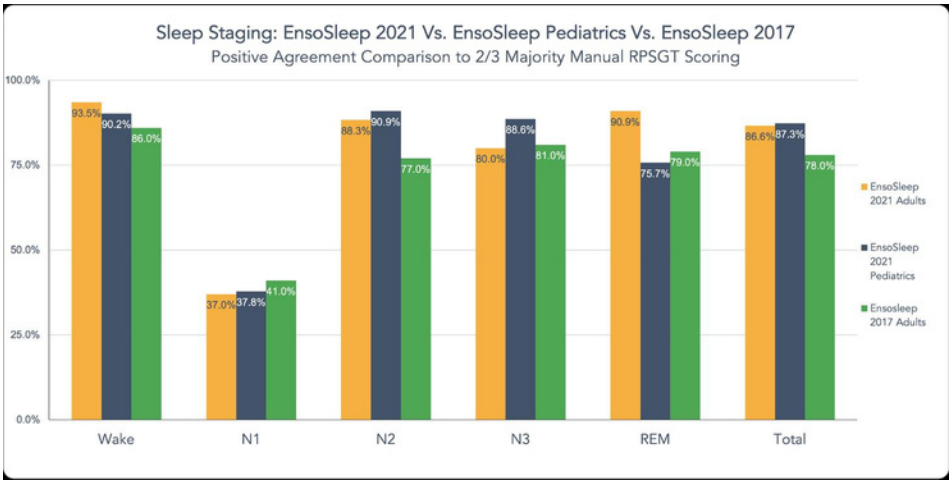
	Two-sided 95% boostrapped median percentile confidence interval (R=2000 resamples)			
	EnsoSleep PPG-TST vs. RR Sample	EnsoSleep EEG-TST vs. RR Sample	EnsoSleep EEG-TST vs. Adult Sample	EnsoSleep EEG-TST vs. Pediatric Sample
Deming Regression Slope B1	0.964 (0.860, 1.067)	0.984 (0.925, 1.023)	1.037 (0.974, 1.201)	1.007 (0.860, 1.067)
Deming Regression Intercept B0 (hrs)	0.089 (-0.484, 0.663)	0.156 (-0.071, 0.504)	-0.181 (-1.101, 0.182)	0.050 (-0.073, 0.226)
Bland-Altman mean difference (MD) [min]	5.380 (2.372, 8.475)	-4.785 (-6.131, -2.237)	0.515 (-4.173, 2.331)	-5.365 (-6.233, -4.015)
Bland-Altman upper limit (ULOA) 95% [min]	73.463 (68.332, 78.743)	32.922 (30.625, 37.269)	57.750 (49.751, 60.849)	17.914 (16.433, 20.217)
Bland-Altman lower limit (LLOA) 95% [min]	-62.703 (-67.835, 57.423)	-42.492 (-44.789, 38.145)	-56.720 (-64.718, 53.621)	-28.644 (-30.126, 26.342)

Table 7. Two-sided 95% bootstrapped confidence intervals for the median percentile for Adult, Pediatric, and Respiratory Rate Samples.

Discussion

Sleep Staging - Adult and Pediatric

EnsoSleep sleep staging event detection agreement performance met all PA, NA, and OA vs 2/3 Majority Scoring acceptance criteria defined when compared to the predicate device (K162627) on a pooled-epochs basis. EnsoSleep staging event detection met the objective pass/fail criteria in both study samples (N=100 Adult Sample and N=100 Pediatric Sample and in all 6 events evaluated (Wake, N1, N2, N3, REM, and Total across all stages).



Graphic 5. Sleep Staging for EnsoSleep 2021 Adult Vs. Pediatrics Vs. EnsoSleep 2017 Adult

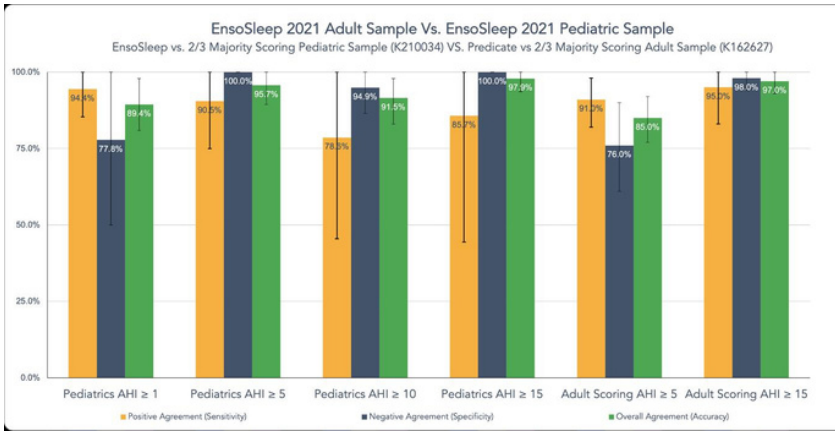
All 3 EnsoSleep PA, NA, and OA point-estimates vs 2/3 Majority Scoring were observed to be greater than the predicate device PA, NA, and OA point-estimates vs 2/3 Majority Scoring in some events in the Adult Sample (Wake, N2, REM, Total) and Pediatric Sample (Wake, N2, N3, Total). Additionally, some of those event detection differences that were in all 3 performance categories (PA/NA/OA) represented a statistically significant result, based on low/upper-bound comparison of two-sided 95% bootstrap percentile method confidence intervals, in each sample respectively; Adult Sample (REM) and Pediatric Sample (N3). None of the 6 events evaluated were observed with PA, NA, or OA point-estimates vs 2/3 Majority Scoring that were 10% or lower in any of the sleep staging event types evaluated. We compare the performance of 2017 EnsoSleep with 2021 EnsoSleep Pediatrics sample below.

Comparison	2021 EnsoSleep (Pediatrics)	2017 EnsoSleep
Wake	94%	86%
N1	37%	41%
N2	88%	77%
N3	80%	81%
REM	91%	79%

Table 8. 2021 Pediatric Sleep Scoring Compared to 2017 EnsoSleep Scoring

Sleep Apnea Diagnostic - Adult and Pediatric

EnsoSleep sleep apnea diagnostic agreement performance met all PA, NA, and OA vs 2/3 Majority Scoring acceptance criteria defined when compared to the predicate device (K162627) vs 2/3 Majority Scoring on a subject-by-subject condition positive or negative agreement basis. EnsoSleep sleep apnea diagnostic agreement met the objective pass/fail criteria in both of the study samples evaluated (N=100 Adult Sample and N=100 Pediatric Sample that included 47 patients in the indicated age range of 13 and older) and in all 4 OSA severity categories assessed (Pediatrics: AHI ≥ 1, AHI ≥ 5, AHI ≥ 10, AHI ≥ 15, Adults AHI ≥ 5, AHI ≥ 15).



Graphic 6. AHI Diagnostics for EnsoSleep 2021 Adult vs. Pediatrics

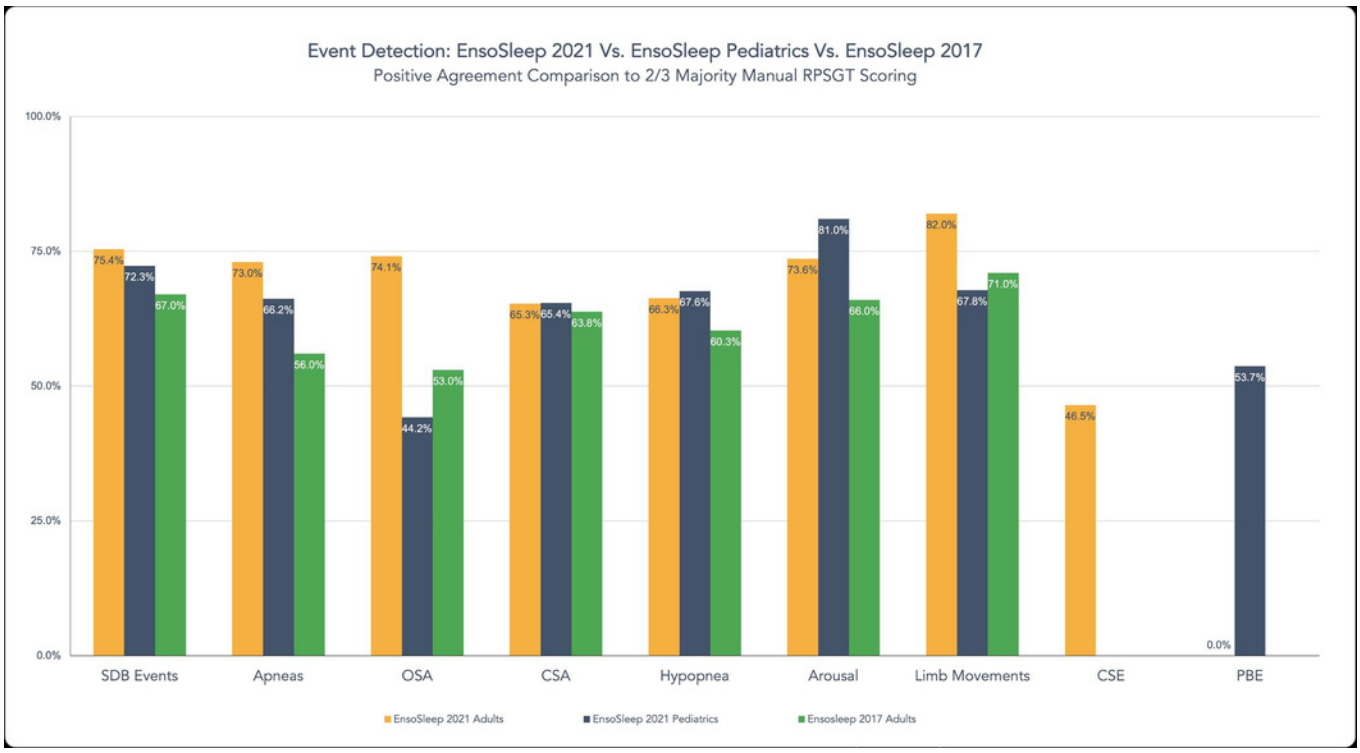
	Likelihood Ratio (+)	Likelihood Ratio (-)
Pediatrics AHI ≥ 1	4.190 (1.892, ∞)	0.070 (0.000, 0.205)
Pediatrics AHI ≥ 5	∞ (∞, ∞)	0.095 (0.000, 0.250)
Pediatrics AHI ≥ 10	15.692 (5.067, ∞)	0.222 (0.000, 0.578)
Pediatrics AHI ≥ 15	∞ (∞, ∞)	0.143 (0.000, 0.556)
Adults AHI ≥ 5	3.76	0.12
Adults AHI ≥ 5	52.25	0.05

Table 9. Corresponding likelihood ratios for each of the six AHI thresholds in Graphic 6.

All three EnsoSleep PA, NA, and OA point-estimates vs 2/3 Majority Scoring were observed to be greater than the predicate device PA, NA, and OA point-estimates vs 2/3 Majority Scoring in some OSA severity categories in the Adult Sample (AHI ≥ 5) and Pediatric Sample (AHI ≥ 5). There were no samples or OSA severities for which there were statistically significant differences observed in all three performance measures (PA/NA/OA), based on low/upper-bound comparison of two-sided 95% bootstrap percentile method confidence intervals. None of the four OSA severity categories evaluated were observed with PA, NA, or OA point-estimates vs 2/3 Majority Scoring that were 10% or lower in any of the three diagnostic agreement performance criteria evaluated (PA/NA/OA) respectively.

Sleep Event Detection - Adult and Pediatric

EnsoSleep scoring event detection agreement performance met all PA, NA, and OA vs 2/3 Majority Scoring acceptance criteria defined when compared to the predicate device (K162627), or reference predicate device (K112102) limited only to hypopnea and central sleep apnea events, on a pooled-epochs basis. Scoring detection met the objective pass/fail criteria in both of the study samples (N=100 Adult Sample and N=100 Pediatric Sample) and in 11 events evaluated (SDB Events, Hypopneas, Apnea, OSA, CSA, Arousal, Limb Movement, RERA, CSE, PBE).



Graphic 7. Event Detection for 2021 EnsoSleep Adults Vs. Pediatrics Vs. 2017 EnsoSleep Adults

All three EnsoSleep PA, NA, and OA point-estimates vs 2/3 Majority Scoring were observed to be greater than the designated predicate device PA, NA, and OA point-estimates vs 2/3 Majority Scoring in some events in the Adult Sample (SDB, Hypopnea, OSA, CSA, Arousal) and the Pediatric Sample (SDB, Hypopnea, CSA, Arousal, RERA). Additionally, some of those event detection differences that were in all three performance categories (PA/NA/OA) represented a statistically significant result, based on low/upper-bound comparison of two-sided 95% bootstrap percentile method confidence intervals, in each sample respectively; Adult Sample (SDB, OSA, and Arousal), and Pediatric Sample (SDB, Hypopnea, Arousal, RERA). None of the 11 events evaluated were observed with PA, NA, or OA point-estimates vs 2/3 Majority Scoring that were 10% or lower in any of the scoring event types evaluated. EnsoSleep clinical validation results for scoring event detection performance met all PA, NA, and OA acceptance criteria in each sample and in each event type evaluated.

Conclusion

The recent AASM position paper on AI in sleep states that “sleep medicine is well positioned to benefit from advances that use big data to create artificially intelligent computer programs. One obvious initial application in the sleep disorders center is the assisted (or enhanced) scoring of sleep and associated events during polysomnography (PSG).” [4] This study validated the ability of AI to achieve a new level of performance for AI-assisted sleep staging, diagnostic, and event detection.

In summary, EnsoSleep staging, sleep diagnostic, and sleep event detection performance met or exceeded the objective assessment criteria. Our comparison to a double-blind, prospective 2/3 Majority Scoring panel consensus reference of independent, qualified sleep technologist scorers, provides additional objective evidence that EnsoSleep is safe and effective for the device indications for use.

Our performance targets were set to exceed that of EnsoSleep when cleared by the FDA in 2017. We generally met or exceeded agreement benchmarks based on previously published inter-scorer reliability standards [2,3].

Key strengths of the current study include:
- Semi-prospective, double-blind, cross-sectional study design; including a large, demographically diverse sample population and controlled for all relevant disease condition, medication, and other relevant confounds, which supports the generalizability of algorithm performance
- Selection criteria ensured that the performance of EnsoSleep was assessed across the full range of disease severity
- Double-blinded, 2/3 Majority Comparative Reference represents a new standard for manual scoring comparison, especially important given variability across scorers [2,3]

EnsoSleep Waveform AI scoring is a validated tool, compatible with most PSG and HSAT hardware/software platforms, that can be implemented with limited or no impact on existing clinician workflows. The consistency of an AI-assisted scoring process can be beneficial in both clinical and research settings by minimizing inter- and intra-scoring variability and providing operational efficiencies. The implementation of EnsoSleep has the potential to free up clinical staff to perform other duties and improve the end-to-end sleep care experience. To learn more about our real world evidence, please refer to the **EnsoData website** to view our **case studies**.

Acknowledgements

Of the many important and impactful contributors to this research, one that stands above for specific acknowledgment is Dr. Nathaniel Watson, MD, of the University of Washington.

Abbreviation Guide

- AI - Artificial Intelligence
- ML - Machine Learning
- SDB - Sleep Disordered Breathing
- HYP - Hypopneas
- OSA - Obstructive Sleep Apnea
- CSA - Central Sleep Apnea
- CSE - Cheyne-Stokes Respiration Episode
- PBE - Periodic Breathing Episode

References

1. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. BMJ. 2004;329(7458):168-169. doi:10.1136/bmj.329.7458.168

2. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. J Clin Sleep Med 2013;9(1):81-87.

3. Rosenberg RS, Van Hout S. The American Academy of Sleep Medicine Inter-scorer Reliability program: respiratory events. J Clin Sleep Med 2014;10(4):447-454.

4. Goldstein CA, Berry RB, Kent DT, et al. Artificial intelligence in sleep medicine: an American Academy of Sleep Medicine position statement. J Clin Sleep Med. 2020;16(4):605-607. doi:10.5664/jcsm.8288