# Clinical Validation of AI Scoring in Adult and Pediatric Clinical PSG Samples Compared to Prospective, Double-Blind Scoring Panel


Chris R. Fernandez, MS[1], Sam Rusk, BS[1], Nick Glattard, MS[1] , Yoav N. Nygate, MS[1], Fred Turkington, BS[1] , Nathaniel F. Watson, MD, MS[2]

[1]EnsoData Research, EnsoData, Madison, WI, USA
[2]Department of Neurology, University of Washington School of Medicine, Seattle, WA, USA


## Introduction

Despite an appreciable rise in sleep wellness and sleep medicine artificial intelligence (AI) research publications, public data corpuses, institutional support, and health AI research funding opportunities, the availability of controlled-retrospective, hybrid-retrospective-prospective, and prospective randomized control trial (RCT) quality clinical validation study evidence is limited with respect to their potential clinical impact. Furthermore, only a few practical examples of AI technologies are validated, in use today clinically, and widely adopted, to assist in sleep diagnoses and treatment. In this study, we contribute to this growing body of clinical AI validation evidence and experimental design methodologies with an interoperable AI scoring engine in adult and pediatric populations.

## Methodology

### Data Sampling Procedure

- An archived collection of retrospective diagnostic clinical polysomnography (PSG) data was randomly sampled with proportionate allocation across each sleep apnea disease severity quantile.
- N=100 adult subjects were sampled from the archived collection with an apnea-hypopnea index (AHI) mean: 22.4 [95% confidence interval (CI): 18.6%, 26.1%], standard deviation (std): 18.8, median: 17.2, min: 0, max: 109.2.
- N=100 pediatric subjects were sampled from the archived collection with an AHI mean: 8.9 [95% CI: 6.8%, 11.0%], std: 10.7, median: 4.9, min: 0, max: 60.3.
- The adult and pediatric samples were observed and verified to have no statistically significant differences in the sleep apnea disease state distributional characteristics relative to the archived collection population.
- All study samples were verified to contain subjects from all sleep apnea disease state severity quantiles.
- All study samples were verified to contain subjects from all predetermined relevant and/or confounding medical conditions, medications, and demographic groups of interest.
- Based on these verification and control procedures, the study sample was determined to be a representative sample.

### Data Annotation Procedure

- A total of three registered polysomnographic technologists (RPSGTs) were used to score the adult and pediatric samples in order to construct a valid 2/3 Majority Scoring consensus reference.
- The manual scoring of the following event types was performed by the three RPSGTs: Sleep stages, Hypopnea events, Obstructive sleep apnea (OSA) events, Central sleep apnea (CSA) events, Arousal events, Limb movement events, Respiratory effort related arousal (RERA) events, Cheyne-stokes (CS) respiration events, Periodic breathing (PB) events.

### Event Detection Agreement Evaluation Procedure

- To construct the 2/3 Majority scoring, a manual scoring reference is derived for each patient and every event type analyzed by computing the 30 second epochs of which at least two raters agreed on the presence of a given event type.
- The output of this procedure is one vector per-patient, per event type, with each element of the vector representing a single 30 second epoch containing a "1" indicating the presence of a given event type, or "0" indicating the absence of a given event type.
- The AI scoring is translated into the same epoch-by-epoch patient-vector format.
- The two vectors are then compared to one another for the computation of positive agreement (PA), negative agreement (NA), and overall agreement (OA), with a two-sided 95% bootstrap confidence interval (CI) for each metric.

### Sleep Apnea Diagnostic Agreement Evaluation Procedure

- To construct the 2/3 Majority sleep apnea consensus, for each rater, the overall AHI and REM AHI (only for the adult population) are calculated.
- Then, the overall AHI and REM AHI are computed from the AI scoring following the same calculation procedure used to calculate the overall AHI and REM AHI for each rater.
- The 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 for the adult population, as well as AHI ≥ 1 and AHI ≥ 10 for the pediatric population.
- For each rater, "1" is assigned if AHI ≥ 1/5/10/15, and a "0" otherwise.
- The AI scoring's overall AHI and REM AHI are translated into the corresponding qualitative result utilizing the same framework as for the raters at the same predefined diagnostic thresholds
- PA, NA, OA, positive likelihood ratio, and negative likelihood ratio, with a two-sided 95% bootstrap CI for each metric are then calculated in a per-patient format.

## Results: Sleep Staging Agreement

**Table 1. Sleep staging agreement for the adult population.**

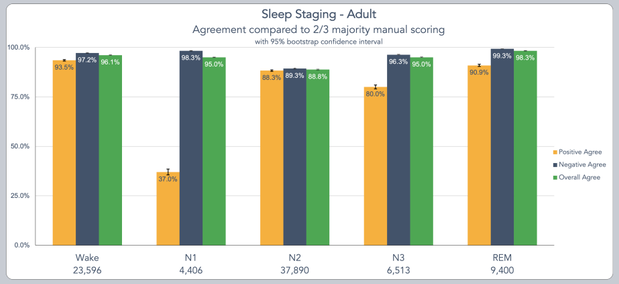
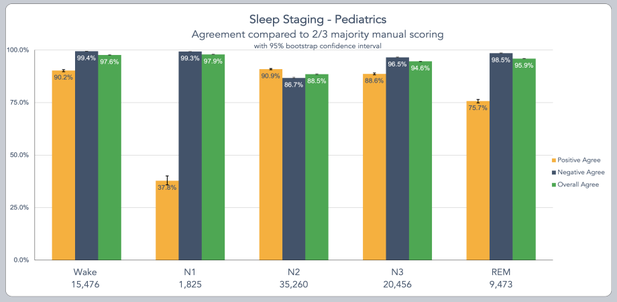
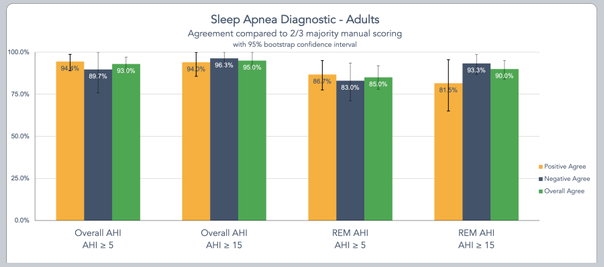



**Table 2. Sleep staging agreement for the pediatric population.**



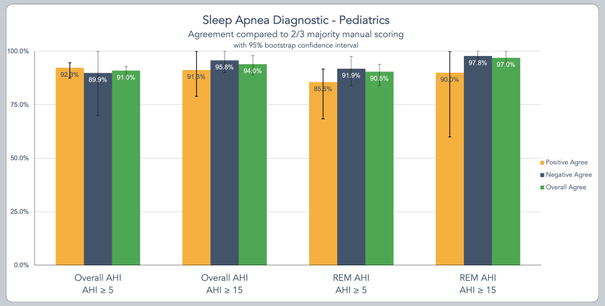

## Results: Sleep Apnea Diagnostic Agreement

**Table 3. Sleep apnea diagnostic agreement for the adult population.**




| 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. Sleep apnea diagnostic agreement for the pediatric population.**




| Likelihood Ratio | Overall AHI AHI > 5 | Overall AHI AHI > 15 | REM AHI AHI > 5 | REM AHI AHI >15 |
|---|---|---|---|---|
| Likelihood Ratio (+) | 9.001 | 22.258 | 10.061 | 45 |
| 95% bootstrap CI | 3.083, ∞ | 9.208, ∞ | 5.258, 31.500 | 15.667, ∞ |
| Likelihood Ratio (-) | 0.097 | 0.09 | 0.163 | 0.102 |
| 95% bootstrap CI | 0.063, 0.140 | 0.000, 0.220 | 0.092, 0.346 | 0.000, 0.400 |

## Sleep Scoring Event Detection Agreement

**Table 5. Sleep scoring event detection agreement for the adult population.**

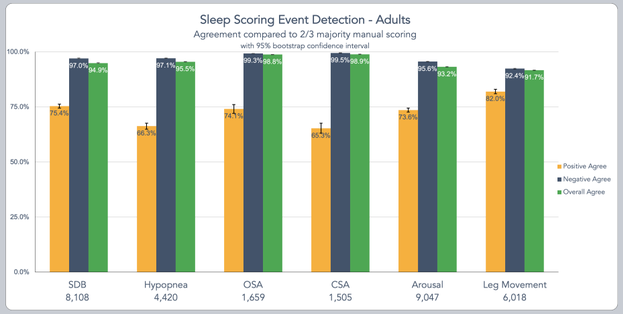



**Table 6. Sleep scoring event detection agreement for the pediatric population.**


## Conclusion

In this study we have provided clinical validation evidence that demonstrates high interoperable AI scoring performance based on retrospective diagnostic clinical PSG recordings of adult and pediatric populations when compared to a prospective, double-blind scoring panel.

Both populations have shown relatively comparable performances for the classification of the five sleep stages, the assignment into two main sleep apnea severity groups, and the detection of six different sleep scoring events.