# Clinical Validation of Artificial Intelligence (AI) Analysis of Photoplethysmogram (PPG) Based Sleep-Wake Staging, Total Sleep Time, and Respiratory Rate

Nathaniel F. Watson, MD, MSc[1], Chris Fernandez, MS[2], Sam Rusk, BS[2], Yoav Nygate, MS[2], Nick Glattard, MS[2], Fred Turkington, BS[2], Justin Mortara, PhD[2]

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

## Introduction

The Photoplethysmogram (PPG) raw waveform is the basis for both the pulse rate and oximetry during polysomnography (PSG) and Home Sleep Apnea Tests (HSAT). The PPG has also recently become ubiquitous as a basis of continuous measurement for the most widely adopted consumer sleep technologies, particularly wearables. In this study, we clinically validate Artificial Intelligence (AI) performance for interoperable, PPG-based epoch-by-epoch Sleep-Wake staging (PPG-S/W), Total Sleep Time (PPG-TST), and Respiratory Rate (PPG-RR), when compared to 1) PSG-based panel 2/3 consensus scoring by registered PSG technologists (RPSGTs) and 2) PSG-based AI scoring of the electroencephalogram (EEG) and respiratory effort belt (EEG-S/W, EEG-TST, Effort Belt-RR).

## Methodology

We applied stratified random sampling with proportionate allocation to a database of N>10,000 retrospective PSGs. We controlled for:

1) Obstructive sleep apnea severity
2) Sleepiness
3) Medical diagnoses including sleep, psychiatric, neurologic, neurodevelopmental, cardiac, pulmonary, and metabolic disorders
4) Medications including benzodiazepines, antidepressants, stimulants, opiates, sedative-hypnotics
5) Demographics including sex, age, BMI, and height and weight

In this manner we established representative adult (N=100) PSG studies from which PPG samples were obtained. Double blinded scoring was prospectively collected for each PSG by 3 experienced RPSGTs randomized from a pool of 6 scorers. RR was established by mode when two scorers agreed on RR value and median otherwise.

## Results

AI EEG-S/W demonstrated 96%/94%/95% Sensitivity, Specificity, and Accuracy compared to RPSGT 2/3 consensus PSG staging. AI PPG-S/W demonstrated 90%/89%/90% Sensitivity, Specificity, and Accuracy compared to the same PSG panel. AI EEG-TST achieved a Pearson Correlation Coefficient (R-value) of 0.968 and AI PPG-TST achieved 0.873 R-value compared to RPSGT 2/3 consensus PSG-TST. When compared to the RR panel consensus in N=282 one-minute RR scoring epochs of PSG, AI Effort Belt-RR performance was <= 2 breaths-per-minute (brpm) in 93.6% of epochs with an average difference of 0.992 brpm, and AI PPG-RR performance was <= 2 brpm in 92.2% of epochs with an average difference of 0.996 brpm.

**Figure 1.** Deming regression scatter plots A) comparing RPSGT 2/3 consensus total sleep time (TST) scoring to AI PPG-TST, and B) comparing RPSGT 2/3 consensus-TST scoring to AI EEG-TST. Overall the results are comparable. The AI PPG-TST slightly overestimates sleep compared to RPSGT 2/3 consensus-TST at lower TST, and then crosses over to slightly overestimate total sleep time compared to RPSGT 2/3 consensus-TST at higher TST. The AI EEG-TST slightly underestimates TST as compared to RPSGT 2/3 consensus-TST initially, then as TST increases the AI EEG-TST and RPSGT 2/3 consensus-TST become increasingly similar.

A. B.

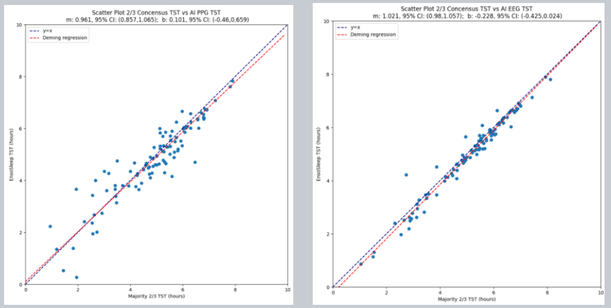



**Figure 2.** Bland-Altman plot showing A) the difference between RPSGT 2/3 consensus-TST and AI PPG-TST (y-axis) plotted vs. RPSGT 2/3 consensus-TST (x-axis) and B) the difference between RPSGT 2/3 consensus-TST and AI EEG-TST (y-axis) plotted vs. RPSGT 2/3 consensus-TST (x-axis). Generally speaking, the mean difference was less for the AI PPG-TST measure, while the limits of agreement were narrower for the AI EEG-TST as compared to RPSGT 2/3 consensus-TST. All told, AI PPG-TST performed favorably to AI EEG-TST when compared to RPSGT 2/3 consensus-TST.

A. B.

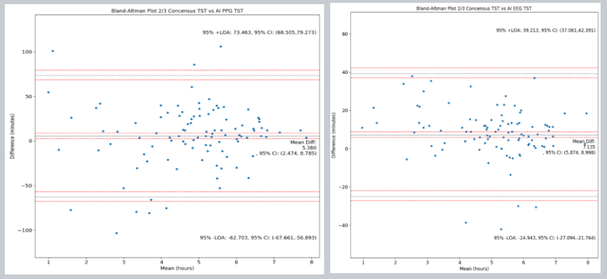



**Figure 3.** Confusion matrices of epoch by epoch sleep/wake (S/W) determination comparing A) AI PPG-S/W vs. RPSGT 2/3 consensus-S/W and B) AI EEG-S/W vs. RPSGT 2/3 consensus-S/W. Overall, AI PPG-S/W agreed 90%, and AI-EEG-S/W agreed 96%, with RPSGT 2/3 consensus-S/W. For epochs that were incorrect when compared to RPSGT 2/3 consensus-S/W, AI PPG-S/W was biased to call sleep epochs wake, and AI EEG-S/W was biased to call wake epochs sleep.

A. B.

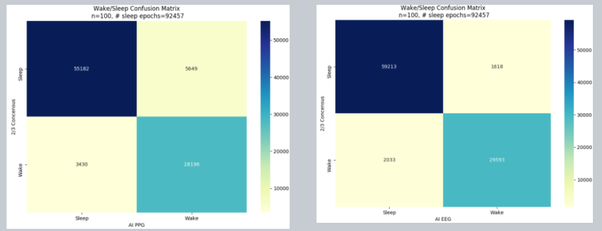
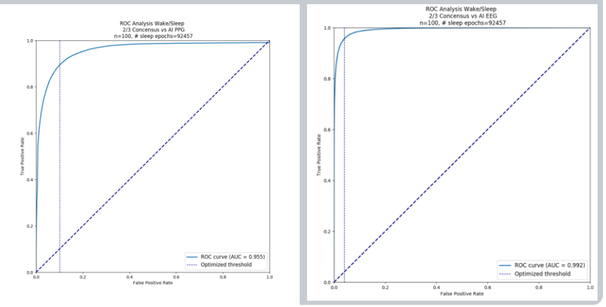



**Figure 4.** Receiver operating characteristic (ROC) curves comparing sleep/wake (S/W) differentiation for A) AI PPG-S/W vs. RPSGT 2/3 consensus-S/W and B) AI EEG-S/W vs RPSGT 2/3 consensus-S/W from over 92,000 epochs of sleep from 100 patients. The AUC was robust for both AI PPG-S/W (0.96) and AI EEG-S/W (0.99) indicating both methods are highly accurate at differentiating sleep from wake.

A. B.




## Conclusion

The study shows interoperable AI analysis performs robustly in evaluating PPG-based epoch-by-epoch sleep-wake stages, total sleep time, and respiratory rate. This demonstrates state-of-art accuracy when compared to a prospective, double-blinded RPSGT PSG scoring panel. AI PPG also performed comparably to AI EEG when compared to RPSGT PSG scoring. This work has implications for consumer sleep technology sleep measurement accuracy, home sleep apnea testing (HSAT) accuracy (via accurate total sleep time determination), and inpatient sleep monitoring. Increasing HSAT accuracy and reliability and reducing indeterminate HSAT results has potential to increase access to sleep care for millions of individuals across the globe.