An unbiased, efficient sleep-wake detection algorithm for a population with sleep disorders: Change Point Decoder

Ayse S. Cakmak[1], Giulia Da Poian[2], Adam Willats[3], Ammer Haffar[4], Rami Abdulbaki[4], Yi-An Ko[5], Amit J. Shah[6,7], Viola Vaccarino[6,7], Donald L. Bliwise[8], Christopher Rozell[1],

Gari D. Clifford[2,3]

[1] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

[2] Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, USA

[3] Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, USA

[4] Rollins School of Public Health, Emory University, Atlanta, GA, USA

[5] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA

[6] Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA

[7] Department of Medicine, Emory University School of Medicine, Atlanta, GA, USA

[8] Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA

* This work was performed in Emory University, Department of Biomedical Informatics

Corresponding author: Ayse Selin Cakmak

Author address: 756 W Peachtree St NW, Floor 15, Atlanta, GA 30308

Author e-mail: acakmak3@gatech.edu

Abstract

Study Objectives: The usage of wrist-worn wearables to detect sleep-wake states remains a formidable challenge, particularly among individuals with disordered sleep. We developed a novel and unbiased data-driven method for detection of sleep-wake and compared its performance to the well-established Oakley algorithm (OA) relative to polysomnography (PSG) in elderly men with disordered sleep.

Methods: Overnight in-lab PSG from 102 participants was compared to accelerometry and photoplethysmography simultaneously collected with a wearable device (Empatica E4). A binary segmentation algorithm was used to detect change points in these signals. A model that estimates sleep or wake states given the changes in these signals was established (Change Point Decoder, CPD). The CPD's performance was compared to the performance of the OA in relation to PSG.

Results: On the Testing Set, OA provided sleep accuracy of 0.85, wake accuracy of 0.54, AUC of 0.67, and Kappa of 0.39. Comparable values for CPD were 0.70, 0.74, 0.78, and 0.40. The CPD method had sleep onset latency error of -22.9 minutes, sleep efficiency error of 2.09%, and underestimated the number of sleep/wake transitions with an error of 64.4. The OA method's performance was 28.6 minutes, -0.03%, and -17.2 respectively.

Conclusions: The CPD aggregates information from both cardiac and motion signals for state determination as well as the cross-dimensional influences from these domains. Therefore, CPD classification achieved balanced performance and higher AUC, despite underestimating sleep/wake transitions. The CPD could be used as an alternate framework to investigate sleep/wake dynamics within the conventional time frame of 30-second epochs.

Keywords: Sleep/wake, change point detection, wearable device, actigraphy, heart rate

Statement of Significance: Wearable devices enabled collecting various physiological signals and sleep assessment for non-laboratory settings. However, most of the proposed methods for sleep/wake detection with wearable devices aim for high overall accuracy at the expense of wake detection performance. The Change Point Decoder (CPD) technique is a novel signal processing approach that can distinguish wakefulness from sleep by solely using changes in the signals collected by wearables. The technique uses temporal information in the changes and the coupling between multiple sources to optimize classification. The results suggest that CPD provides unbiased sleep/wake detection with performance comparable to a traditional algorithm for sleep efficiency but with potential underestimation of sleep/wake transitions.

Introduction

Several sleep/wake classification algorithms for wearables have been suggested over the last decades, and they are typically based solely on actigraphy derived from accelerometer.[1-4] Several findings suggest that only using movement signals leads to the main limitation of current algorithms: the incorrect classification and overestimation of low activity tasks as such sleep.[5-7] Indeed, low activity (quiescent) segments are not unique to sleep but are common to other activities such as reading or watching television. Another limitation results from the adoption of imprecise evaluation metrics used in assessing the performance of these devices. Since the percentage of sleep is typically higher compared to wake overnight, total accuracy may not be a reliable metric to evaluate performance. Sleep/wake detection may be considered as a "rare class problem" and may be amenable to alternative model evaluation metrics which better reflect this issue.

The first approaches in the field for state determination were based on calculating a weighted sum over the actigraphy epochs around the current epoch and scaling the summation to distinguish sleep from wakefulness.[1,2] Oakley presented a similar approach in which the current epoch, epochs in the 2 minutes before and the 2 minutes after the current epoch are scaled with predetermined coefficients and summed.[8] If the summation is higher than the threshold, the region was labeled as wake. The Oakley algorithm is utilized in commercially available devices with different threshold selections (e.g., Actiwatch 2, Philips Respironics; Bend, Oregon). These actigraphic methods rely solely on the amplitude of actigraphic signals, which makes them low cost and easy to implement. However, these methods may overestimate sleep, particularly for patients with disordered sleep.[9,10] It has been long been known that heart rate reflects transitions from sleep to wake and from wake to sleep.[11-13] Recent studies in the field leverage a combination of photoplethysmography (PPG) and accelerometer signals for sleep/wake detection.[14-16] However, these approaches have not been tested on clinical populations and still show low sensitivity in detecting wake epochs.

The method described in this paper combines PPG and accelerometer signals collected from wearable devices and detects patterns in change points associated with sleep/wake transitions. The proposed method, which is referred to as Change Point Decoder (CPD), is inspired by methods related to neural spike train models and uses a similar encoding/decoding framework.[17] In this study, CPD was developed on a clinical data set of 102 patients, which was split into a training set of 70 subjects and a test set of 32 subjects. The effect of sleep disorders on the method's performance was then investigated.

Methods

Participants

The current study includes a subgroup of participants (n = 102, men, mean age = 68.56, SD = 1.93) from the Emory Twin Study Follow-up recruited from the Vietnam Era Twin Registry.[18] All Polysomnography (PSG) data were collected from data acquisition systems (Natus, Remlogic) set up in two bedrooms in the Emory Sleep Center. Written informed consent was obtained from all participants, and the Emory University Institutional Review Board approved this research (IRB # 00081004). During PSG acquisition, subjects wore a commercially available wrist-worn watch (Empatica E4, Empatica; Cambridge, MA). The wrist-worn device recorded Photoplethysmogram (PPG) and three-axis accelerometer signals.

Data Set

The study population was assigned to four groups according to their Apnea-Hypopnea Index (AHI) and Periodic Limb Movement Index (PLMI) as follows:

- **Group 1:** Subjects with AHI < 15 and PLMI < 15
- **Group 2:** Subjects with AHI ≥15 and PLMI < 15
- **Group 3:** Subjects with AHI < 15 and PLMI ≥15
- **Group 4:** Subjects with AHI ≥15 and PLMI ≥15

All the data were randomly split into two sets, with 70 subjects assigned to the training set and 32 subjects assigned to testing. Table 1 show ages and PSG-defined sleep efficiency in both sets. Two-sample Kolmogorov tests were performed for age, AHI, PLMI, and sleep efficiency of the subjects in the training and testing sets. Differences in these measures between the sets were not statistically significant, suggesting that the training set is representative of the testing set.

Preprocessing of Signals

Previous studies have shown that the mean and standard deviation of heart rate decreases during Non-REM sleep and increase during wakefulness.[11,13] We hypothesized that change point detection could be used to mark these alterations in the heart rate. Body movements have also been used as a sleep/wake identification feature in various studies over the years.[1,7,8] In this study, changes in the amplitude and gross body movements were detected to capture this information.

Initially, the Empatica E4 timestamp was synchronized with the PSG timestamp. The next preprocessing step consisted of converting the PPG signal to Normal-to-Normal (NN) beat interval time series and three-axis accelerometer data to actigraphy and angle time series. PPG data were preprocessed using PhysioNet Cardiovascular Signal Toolbox.[19] Firstly, peak detection was performed using the *qppg* method provided with the toolbox, and the data was converted to peak-to-peak (PP) interval time series. Then, non-sinus intervals were detected and removed by measuring the change in the current PP interval from the previous PP interval and excluding intervals that change by more than 20%. PP intervals outside of physiologically possible range were also removed to obtain NN interval time series, which was filtered using a Kalman filter to reduce noise.[20,21]

Raw three-axis accelerometer data were converted to activity counts following the approach described by Borazio *et al.*[22] Activity counts are the output format of most commercial actigraphy devices; data are summarized over 30-second epochs or time intervals. This conversion compresses information, reduces required memory for storing data, and eliminates artifacts and noise in raw

data. Z-axis actigraphy data were filtered using a 0.25-11 Hz passband to eliminate extremely slow or fast movements.[23] The maximum values inside 1-second windows were summed for each 30-second epoch of data to obtain the activity count for each epoch.

Lastly, a tilt angle time series was derived from the raw accelerometer data to capture information that is not present in the activity count time series. Specifically, tilt angle, which is the angle between the gravitational vector measured by the accelerometer and the initial orientation with the gravitational field pointing downwards along the z-axis, can be calculated from the accelerometer reading as

$$\rho = \frac{a_z}{\sqrt{a_x^2 + a_y^2 + a_z^2}}, \quad (1)$$

where ρ is the tilt angle and a_x, a_y, and a_z are the readings from x, y, and z axes of the accelerometer respectively.

Change Point Detection

Binary Segmentation (BiS) was used on the preprocessed actigraphy, tilt angle, and NN interval time series to detect significant changes in the mean and standard deviation, as seen in Figure 1. BiS technique was chosen for the its simplicity and easy implementation. The procedure starts by searching for a change point τ in the input signal $S = \{x_1, x_2, \dots, x_N\}$ that satisfies the condition

$$C_{S_{1:\tau}} + C_{S_{\tau+1:N}} + \beta < C_{S_{1:N}} \quad (2)$$

where C is a cost function and β is a penalty term that reduces overfitting. If the condition in Eq. 2 is met, τ becomes the first estimated change point, and $S_{1:\tau}$ and $S_{\tau+1:N}$ become the first subsequences. The process continues within these subsequences until data cannot be divided any further. Cost function in the above equation is given by

$$C_{S_{\tau_{i-1}:\tau_i}} = -2\log\mathcal{L}\left(\theta\middle|S_{\tau_{i-1}:\tau_i}\right) \quad (3)$$

where $\mathcal{L}$ is the likelihood function.

In a previous study, Yoneyama *et al.* selected body movements with more than 10° changes in the body angle as turnover events.[24] They used the bi-modal distribution of turnover angle changes and duration between turnovers to analyze sleep in healthy and neurodegenerative patients. Those authors also stated that abdominal motion due to breathing causes 5° fluctuations, so 10° threshold is ideal for detecting turnover events. In this study, changes more than 10° tilt angle were used as a change point in the tilt angle time series.

Encoding Generalized Linear Models

The CPD model is inspired by the encoding/decoding framework in neuroscience,[17] where a neural population response to a stimulus signal is observed in the form of spike trains. These responses are used to train encoding models which describe the probability of the responses. Then, when a spike train is observed from a group of cells, this model is used to "decode" or estimate the stimulus signal. Similarly, in the proposed CPD model, the sleep/wake signal through the night was thought as the stimulus driving the changes in the NN time series and actigraphy signals collected by the wearable device. Following the approach by Pillow *et al*, the information in the change point time series was used to train the encoding model. As seen in Figure 2, the model consists of a history filter, coupling filters, and a stimulus filter. In the encoding step, the optimal filters are selected using the training data. For example, the instantaneous firing rate of NN time series can be expressed as

$$r_{NN}(t) = f\left(k_{NN}\cdot x(t) + h\cdot z_{NN,history}(t) + c_{NN,act}\cdot z_{act}(t) + c_{NN,angle}\cdot z_{angle}(t)\right) \quad (4)$$

where $x(t)$ is the sleep/wake stimulus that drives the changes in the signals. k, h and c are stimulus, history, and coupling filters respectively. $z_{NN,history}$ represents the history of the NN time series while z_{act} and z_{angle} are the windows of actigraphy and angle time series. f can be selected as the exponential function and it converts the summation into probability of spiking. We fitted this set of

four filters for each actigraphy, angle, and NN time series. Filter coefficients were calculated by using "glmfit" function from MATLAB.[25] This generalized linear model approach allowed for both excitatory and inhibitory interactions between signals.

Decoding Generalized Linear Models

The decoding framework is composed of three steps as shown in Fig. 3. The decoding uses the trained model from the encoding step and tries to estimate if the subject is asleep or wake, given the changes in the input signals. The change point time series derived from each of the three data streams were fed into the trained model, and the penalized maximum likelihood estimate of the sleep/wake stimulus was calculated by minimizing

$$\hat{x} = \underset{x}{\operatorname{argmin}}(-\log p(x|z) + \lambda \|x\|_{TV}) \quad (5)$$

where $\hat{x}$ is the estimate sleep/wake and z is the change point time series and $\log p(x|z)$ is the log-probability of sleep/wake states given the observed change events. We regularized the likelihood with the Total Variation (TV) norm to prevent overfitting and preserve step-like properties of the sleep/wake stimulus. After estimation, the output $\hat{x}$ is thresholded and converted back to binary sleep/wake detection as seen in Fig. 4. More details on encoding and decoding steps can be found in the Supplement section.

The data window size for encoding model filters, the TV regularization parameter λ, and the threshold were selected by sweeping a range of values and selecting parameters maximizing the F1 score in the training set. Data window sizes tested were 30 seconds, 1 minute and 1.5 minutes. Fig. 5 illustrates the sweep of regularization parameter in the range [0:0.1:5] and threshold [0:0.01:0.5]. F1 score was used to guide model selection because it is a combined metric for precision and recall. Precision indicates how many epochs of detected wake are correct, whereas recall refers to the percentage of total wake epochs results correctly classified. Therefore, F1 score, which combines precision and recall, proves to be a useful metric for this imbalanced classification scenario.

Oakley Method

The Oakley sleep/wake detection method was also implemented on the same dataset to allow a fair comparison with the proposed technique. The algorithm is adapted for 30-second epochs following the approach by Kosmadopoulos *et al.*[26] Actigraphy data are weighted and summed as follows

$$A_i = 0.04\,E_{(i-4)} + 0.04\,E_{(i-3)} + 0.2\,E_{(i-2)} + 0.2\,E_{(i-1)} + 2\,E_{(i)} + 0.2\,E_{(i+1)} + 0.2\,E_{(i+2)} + 0.04\,E_{(i+3)} + 0.04\,E_{(i+4)} \quad (7)$$

where i denotes the current epoch index and E denotes the actigraphy count in the epoch. Then A_i is compared to a predefined threshold to identify sleep/wake. In commercially available Actiwatch devices, there are three different thresholds: low (20), medium (40), and large (80). Since the wearable device is different in this study, it could result in an actigraphy time series with a different amplitude range than Actiwatch and thresholds may not apply. Therefore, the threshold was selected using the training data to maximize F1 score. Results of both optimized threshold and medium setting are reported for comparison.

Performance Evaluation

To evaluate the performance of the model, standard metrics such as sleep accuracy, wake accuracy, and total accuracy were calculated. F1 score is used both for hyper-parameter selection as described above and for evaluating the algorithms. Also, we fixed the regularization parameter of CPD to the value selected using the F1 score and sweep thresholds for both methods to derive ROC and Precision-Recall curves. Cohen's Kappa was also calculated to measure inter-rater reliability between PSG study and the algorithms. Furthermore, sleep-wake statistics including Wake After Sleep Onset (WASO), Sleep Onset Latency (SOL), Sleep Efficiency, and the number of sleep wake transitions were calculated. WASO was defined as the minutes awake during the sleep period after sleep onset (defined as the first 30-second epoch of any stage of sleep). Sleep Onset Latency was calculated as the time from lights out until sleep onset in minutes. Sleep efficiency was defined as

the percent of time scored as sleep during the sleep period subsequent to sleep onset. For training set performance evaluation, models were trained and validated using leave-one-out cross validation within training set. For testing set performance evaluation, final model was trained using the subjects in the training set with selected hyperparameters and tested on the testing set. Using individual signal models without the coupling filters between different domains was also tested in the same manner in order to assess the contribution of each signal and the coupling filters to the performance.

Results

Hyperparameters selected on training set for CPD are 1-minute window size, regularization parameter of 2, and threshold of 0.22. For the OA method, threshold optimized with F1 score on the training set is equal to 70. Concordance between PSG and the two methods are evaluated on testing set. The mean across subjects for total accuracy, sleep accuracy, wake accuracy, Kappa, F1 score, WASO, and SE are shown in Table 2 and Table 3 for both methods. For WASO, SE and the number of sleep wake transitions, the error is calculated as the PSG gold standard minus estimated value. Fig. 6 illustrates Receiver Operating Characteristic (ROC) curve and Precision-Recall curve for both methods as their threshold is varied. Operating points selected using the training data are also marked with red circles on the plots. The area under the curve (AUC) for the CPD method was found to be 0.78 and 0.67 for the OA method. Moreover, we observed from Fig. 6 that it was possible to achieve similar performance to OA by changing the CPD method's threshold. However, it was not possible for OA method to reach the CPD's operating point by modifying the threshold value.

As shown in Table 2 and Table 3, the CPD method achieved greater accuracy for wake accuracy, Kappa, and F1 Score for both training and test sets. The difference between wake accuracy

was statistically significant ($P < 0.05$) for the methods in both training and test sets. It can also be seen that OA overestimated WASO while wake accuracy is low. Note that the CPD method exhibited lower WASO error in both analyses. When using the medium threshold setting (40) is used for the OA method, total accuracy was 0.54, sleep accuracy was 0.38, and wake accuracy was 0.81 for the test set. The error in the number of sleep wake transitions in the test set was overestimated as -17.19 (36.13) for the OA algorithm and underestimated as 64.41 (34.80) for the CPD.

Table 4 shows the same experiment repeated by using each signal by itself, without the coupling filters between the different domains. Tilt angle signal model performed better than PPG and actigraphy models in terms of Kappa, F1 score, WASO error, and SE error performance metrics. However, all three single signal models resulted in lower total accuracy, Kappa, F1 score, and higher SE error when compared to the combined model with the coupling filters.

Figure 7 provides the Bland Altman analyses of the differences for SE and WASO for the OA and CPD methods for the Testing set. The modified Bland Altman plot shows that the Oakley method exhibited a bias towards overestimating WASO (see Figure 7, bottom left subplot). These plots also show that both methods exhibited similar performance as measured by SE error.

Tables 5 and 6 compare the results of both methods for all four groups in the test set. The CPD has a higher wake accuracy than the Oakley method in each subject group, while the Oakley method performs slightly better in terms of total accuracy.

Discussion

This article presents a novel method (CPD) for identifying sleep and wake states from movement and physiological signals collected using wearable devices. The method was comprised of three types of filters; stimulus, history, and coupling. Filter coefficients were estimated through a training process and then were used to detect sleep and wake states from change points. Our approach was flexible enough to incorporate various signal modalities and incorporating information

from these results in higher wake detection performance. The CPD approach used a combination of movement-related and physiological signals, making it possible to overcome some of the limitations of previous algorithms based solely on actigraphy. For instance, the results demonstrate that the CPD method does not overestimate sleep and has high wake detection performance. Therefore, the CPD method can provide an unbiased solution to sleep/wake detection. The CPD modeled time series of discrete change events derived from wearable device signals and outputted a score of wakefulness ($\hat{x}$) which can be used to investigate gradual transitions between sleep and wake states within the epochs.

The OA method exhibited a higher sleep accuracy with respect to the CPD approach, which resulted in slightly higher total accuracy for OA since the prevalence of the sleep epochs in the data was relatively higher than the prevalence of wake epochs. By contrast, we observed a significant improvement in wake accuracy by using the CPD. Higher wake accuracy also resulted in lower WASO error for both training and test sets with the CPD. The OA method overestimated WASO and had lower wake detection accuracy, even though the threshold parameter was optimized during training (Table 2, 3). This outcome indicated that the Oakley algorithm misclassified sleep epochs as wake while being unable to recognize true wake epochs. A similar pattern was observed in subjects without any sleep disorder (Group 1) within the test set. This result could be due to the fact that when there is no movement, OA could not estimate that the subject was wake, as exemplified in Figure 8.

Periodic Limb Movement Disorder is characterized by episodes of limb movements during sleep, and these limb movements could bias the actigraphy based method into estimating a subject is awake. For PLMD subjects (Group 3), the CPD method had higher wake accuracy compared to OA, indicated in Table 5. However, this did not lead to significantly lower WASO Error due to the CPD method's lower sleep accuracy in this group, suggesting that limb movements had a similar effect in both methods.

Accurate estimates of WASO could become especially important in monitoring populations with difficulties falling or staying asleep. For example, WASO duration has been used as a diagnostic criterion for insomnia.[27] The OA method is known to have lower performance in detecting wakefulness for insomnia. [9,10] In this study, optimizing the threshold parameter for OA did not yield a significant increase in wake accuracy. Therefore, the CPD method could be more useful in this population due to its higher accuracy in detecting wake epochs and the lower error in WASO. On the other hand, CPD method had a high error for estimating the number of sleep/wake transitions, which should be taken into account while applying the method on the insomnia population.

The proposed method only required the timestamps of the change points. Due to this fact, the CPD approach required less storage space than other methods. In this study, saving raw accelerometer and PPG signals for each subject resulted in 6.91 GB of data. However, if the change points alone were saved, stored data were only 1.3 MB. Using the CPD method reduced the required memory to 0.02% compared to other approaches that need the whole signal for feature extraction or training the models. As a result, the CPD method could result in immense memory (and energy) savings for large populations, applications with more data streams, and studies in which subjects are monitored over long periods.

This study has some limitations. Since the signals were stored as change point time series and raw signals were not saved, the information in signal segments was lost. This could limit the data being used for other applications such as detecting or monitoring disorders like arrhythmia or sleep apnea. Also, it has been observed that the CPD approach has lower wake accuracy in subjects with sleep apnea (Group 2) compared to other groups. Future studies will explore adding a PPG-derived respiration signal to the model to improve performance in subjects with sleep apnea. A second limitation of the CPD is the lower number of sleep/wake transitions. The CPD method employs total variation regularization. While this regularization prevents overfitting and preserves

piecewise constant structure of sleep/wake signal, it results in fewer switches between sleep and wakefulness.

In conclusion, this work presents the Change Point Decoder, which is a novel technique for sleep/wake identification in patients with highly disordered sleep. The CPD provides higher wake detection accuracy when compared to a solely actigraphy-based method. This superior performance could enable more accurate investigation of the vital role of awakenings during the night in various psychological disorders. The CPD method requires low memory in the wearable devices compared to existing methods, and therefore, it could prove beneficial in long-term studies. Moreover, as a method, the CPD has the ability to adapt to different and novel devices and signals beyond the accelerometer.

Acknowledgements

The authors wish to acknowledge the National Science Foundation (NSF) award # 1636933 "BD Spokes: SPOKE: SOUTH: Large-Scale Medical Informatics for Patient Care Coordination and Engagement", and National Institutes of Health (Grant # NIH 5R01HL136205-02), National Heart, Lung, and Blood Institute (Award # K23 HL127251) for their financial support of this research. Supported is also acknowledged from the NSF under award CCF-1409422 and the James S. McDonnell Foundation award # 220020399.

Disclosure Statement

a) Financial Disclosure: Donald L. Bliwise: Consultant for Ferring, Merck, Eisai and Jazz.

b) Non-financial Disclosure: None.

References

1. Cole R, Kripke D, Gruen W, Mullaney D, Gillin J. Automatic Sleep/Wake Identification From Wrist Activity. *Sleep*. 1992;15(5):461-469. doi:10.1093/sleep/15.5.461
2. Webster J, Kripke D, Messin S, Mullaney D, Wyborney G. An Activity-Based Sleep Monitor System for Ambulatory Use. *Sleep*. 1982;5(4):389-399. doi:10.1093/sleep/5.4.389
3. Lotjonen, J., Korhonen, I., Hirvonen, K. Automatic Sleep-Wake and Nap Analysis with a New Wrist Worn Online Activity Monitoring Device Vivago WristCare®. *Sleep*. 2003;26(1):86-90. doi:10.1093/sleep/26.1.86
4. Hedner J, Pillar G, Pittman S, Zou D, Grote L, White D. A Novel Adaptive Wrist Actigraphy Algorithm for Sleep-Wake Assessment in Sleep Apnea Patients. *Sleep*. 2004;27(8):1560-1566. doi:10.1093/sleep/27.8.1560
5. Goldstone A, Baker F, de Zambotti M. Actigraphy in the digital health revolution: still asleep?. *Sleep*. 2018;41(9). doi:10.1093/sleep/zsy120
6. Marino M, Li Y, Rueschman M et al. Measuring Sleep: Accuracy, Sensitivity, and Specificity of Wrist Actigraphy Compared to Polysomnography. *Sleep*. 2013;36(11):1747-1755. doi:10.5665/sleep.3142
7. Paquet J, Kawinska A, Carrier J. Wake Detection Capacity of Actigraphy During Sleep. *Sleep*. 2007;30(10):1362-1369. doi:10.1093/sleep/30.10.1362
8. Oakley N. *Validation With Polysomnography Of The Sleepwatch Sleep/Wake Scoring Algorithm Used By The Actiwatch Activity Monitoring System*. Bend: Mini Mitter, Cambridge Neurotechnology; 1997.

9. Sivertsen B, Omvik S, Havik O et al. A Comparison of Actigraphy and Polysomnography in Older Adults Treated for Chronic Primary Insomnia. *Sleep*. 2006;29(10):1353-1358. doi:10.1093/sleep/29.10.1353
10. Kang S, Kang J, Ko K, Park S, Mariani S, Weng J. Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *J Psychosom Res*. 2017;97:38-44. doi:10.1016/j.jpsychores.2017.03.009
11. Chouchou F, Desseilles M. Heart rate variability: a tool to explore the sleeping brain?. *Front Neurosci*. 2014;8. doi:10.3389/fnins.2014.00402
12. Varoneckas, G., Plauska, K., Kauk, J. J. Components of the heart rhythm power spectrum in wakefulness and individual sleep stages. *International Journal of Psychophysiology*. 1986;4(2):129-141. doi:10.1016/0167-8760(86)90006-1
13. Bonnet M, Arand D. Heart rate variability: sleep stage, time of night, and arousal influences. *Electroencephalogr Clin Neurophysiol*. 1997;102(5):390-396. doi:10.1016/s0921-884x(96)96070-1
14. Fonseca P, Weysen T, Goelema M et al. Validation of Photoplethysmography-Based Sleep Staging Compared With Polysomnography in Healthy Middle-Aged Adults. *Sleep*. 2017;40(7). doi:10.1093/sleep/zsx097
15. Beattie Z, Oyang Y, Statan A et al. Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals. *Physiol Meas*. 2017;38(11):1968-1979. doi:10.1088/1361-6579/aa9047
16. Eyal S, Baharav A. Sleep insights from the finger tip: How photoplethysmography can help quantify sleep. In: *Computing In Cardiology (Cinc)*. ; 2017.
17. Pillow J, Ahmadian Y, Paninski L. Model-Based Decoding, Information Estimation, and Change-Point Detection Techniques for Multineuron Spike Trains. *Neural Comput*. 2011;23(1):1-45. doi:10.1162/neco_a_00058

18. Tsai M, Mori A, Forsberg C et al. The Vietnam Era Twin Registry: A Quarter Century of Progress. *Twin Research and Human Genetics*. 2012;16(1):429-436. doi:10.1017/thg.2012.122
19. Vest A, Da Poian G, Li Q et al. An open source benchmarked toolbox for cardiovascular waveform and interval analysis. *Physiol Meas*. 2018;39(10):105004. doi:10.1088/1361-6579/aae021
20. Welch G, Bishop G. An introduction to the Kalman filter. 1995.
21. Li Q, Mark R, Clifford G. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter. *Physiol Meas*. 2007;29(1):15-32. doi:10.1088/0967-3334/29/1/002
22. Borazio M, Berlin E, Kucukyildiz N, Scholl P, Van Laerhoven K. Towards benchmarked sleep detection with wrist-worn sensing units. In: IEEE; 2014:125-134.
23. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak C. The Role of Actigraphy in the Study of Sleep and Circadian Rhythms. *Sleep*. 2003;26(3):342-392. doi:10.1093/sleep/26.3.342
24. Yoneyama M, Okuma Y, Utsumi H, Terashi H, Mitoma H. Human turnover dynamics during sleep: Statistical behavior and its modeling. *Physical Review E*. 2014;89(3). doi:10.1103/physreve.89.032721
25. *MATLAB*. Natick, Massachusetts: The MathWorks Inc.; 2018.
26. Kosmadopoulos A, Sargent C, Darwent D, Zhou X, Roach G. Alternatives to polysomnography (PSG): A validation of wrist actigraphy and a partial-PSG system. *Behav Res Methods*. 2014;46(4):1032-1041. doi:10.3758/s13428-013-0438-7
27. Lichstein K, Durrence H, Taylor D, Bush A, Riedel B. Quantitative criteria for insomnia. *Behav Res Ther*. 2003;41(4):427-445. doi:10.1016/s0005-7967(02)00023-2

Figure List

- **Figure 1:** Change points detected from various signals are visualized here as dashed lines. Actigraphy time series is shown on figure (a) and tilt angle time series are shown in figure (b). Figure (c) shows NN interval time series in light purple, Kalman filtered NN intervals in darker purple.
- **Figure 2:** Encoding diagram for angle time series. Raw signals from the Empatica devices are converted to change point time series. History, coupling, and stimulus filters were applied on one-minute windows of data, summed and converted to spiking probability of tilt angle time series in time interval δt using instantaneous firing rate $r_{tilt}(t)$.
- **Figure 3:** Pipeline for decoding sleep/wake stages. Detected changed points are fed into the trained model. Output of the model is converted to sleep/wake detection using penalized Poisson Maximum Likelihood estimation (Eq. 5) and thresholding to convert x_{est} to binary sleep/wake predictions. Sleep and wake states are indicated with 'S' and 'W' letters on the figure.
- **Figure 4:** Example of the proposed CPD algorithm. Top figure: output of the decoder $\hat{x}$ (light green solid line), ground truth (i.e. gold-standard polysomnography) (light-blue area), and threshold (red dashed line). Bottom figure: binary sleep/wake output after thresholding (dark green solid line).
- **Figure 5:** Sweeping threshold and regularization parameter. F1 score represents a combination of precision and recall, with high values reflecting better performance

- **Figure 6:** ROC and Precision-Recall curves for the CPD and OA methods. Performance of both methods is illustrated as their threshold varied. Operating points are shown with red circles on the plots.
- **Figure 7:** Modified Bland Altman plots for sleep metrics in Test Set. x axis shows ground truth (i.e. gold standard) PSG metrics and y axis shows the difference between PSG and the estimates. Participants belonging to four subgroups determined by AHI and PLMI are indicated with different symbols.
- **Figure 8:** Comparison of Oakley and Change Point Decoder Methods. Top plot illustrates actigraphy while middle and bottom plots shows OA and CPD estimates respectively.

Table 1- Participant Demographics and PSG Sleep Statistics in Training and Test Sets. Mean (Standard Deviation) of variables in each group.

	Training Set			Testing Set		
	n	Age	PSG Sleep Efficiency	n	Age	PSG Sleep Efficiency
Group 1	19	68.1 (2.31)	75 (13)	9	68.22 (2.64)	74 (14)
Group 2	22	68.0 (2.03)	72 (14)	7	69.43 (1.62)	74 (13)
Group 3	24	68.42 (2.87)	74 (15)	14	68.28 (1.68)	67 (19)
Group 4	5	67.20 (3.42)	65 (16)	2	69 (0)	74 (5)
All Subjects	70	68.11 (2.48)	73 (14)	32	68.56 (1.93)	71 (16)

Table 2- Sleep/wake identification performances in the Training Set

	Training Set			
	OA	95 % CI	CPD	95 % CI
	Mean (SD)		Mean (SD)	
Total Accuracy	0.76 (0.10)	[0.73, 0.78]	0.76 (0.12)	[0.73, 0.79]
Sleep Accuracy	0.82 (0.15)	[0.79, 0.86]	0.78 (0.19)	[0.73, 0.82]
Wake Accuracy	0.61 (0.19)	[0.56, 0.65]	0.72 (0.18) *	[0.67, 0.76]
Kappa	0.41 (0.17)	[0.37, 0.45]	0.46 (0.20)	[0.41, 0.50]
F1 Score	0.59 (0.14)	[0.56, 0.63]	0.64 (0.15) *	[0.60, 0.68]
WASO Error (min.)	-22.82 (69.04)	[-39.28, -6.36]	13.17 (56.84) *	[-0.38, 26.72]
SE Error (%)	3.01 (15.74)	[-0.74, 6.76]	-1.59 (14.18) *	[-4.97, 1.79]
SOL Error (min.)	22.69 (25.72)	[16.56, 28.83]	-10.49 (43.30) *	[-20.81, 0.16]

* Wilcoxon signed-rank comparison of two methods, 5% significance level.

Abbreviations: CI, Confidence Interval; SD, Standard Deviation.

Table 3- Sleep/wake identification performances in the Testing Set

	Testing Set			
	OA	95 % CI	CPD	95 % CI
	Mean (SD)		Mean (SD)	
Total Accuracy	0.76 (0.09)	[0.72, 0.79]	0.72 (0.14)	[0.67, 0.77]
Sleep Accuracy	0.85 (0.12) *	[0.80, 0.89]	0.70 (0.19)	[0.63, 0.76]
Wake Accuracy	0.54 (0.20)	[0.47, 0.62]	0.74 (0.20) *	[0.66, 0.81]
Kappa	0.39 (0.17)	[0.33, 0.45]	0.40 (0.24)	[0.31, 0.49]
F1 Score	0.59 (0.14)	[0.54, 0.64]	0.62 (0.20)	[0.55, 0.70]
WASO Error (min.)	-9.95 (63.75)	[-32.94, 13.03]	7.66 (67.34)	[-16.62, 31.94]
SE Error (%)	-0.03 (14.93)	[-5.42, 5.35]	2.09 (16.81)	[-3.97, 8.15]
SOL Error (min.)	28.64 (36.84)	[15.36, 41.92]	-22.86 (58.68) *	[-44.01, -1.7]

* Wilcoxon signed-rank comparison of two methods, 5% significance level.

Abbreviations: CI, Confidence Interval; SD, Standard Deviation.

Table 4- Performances of single signal models in the Testing Set

	PPG model Mean (SD)	Act. model Mean (SD)	Tilt model Mean (SD)
Total Accuracy	0.60 (0.14)	0.69 (0.13)	0.69 (0.15)
Sleep Accuracy	0.49 (0.19)	0.89 (0.12)	0.65 (0.22)
Wake Accuracy	0.83 (0.13)	0.34 (0.17)	0.75 (0.20)
Kappa	0.25 (0.17)	0.24 (0.19)	0.36 (0.23)
F1 Score	0.60 (0.16)	0.41 (0.17)	0.61 (0.19)
WASO Error (min.)	-53.31 (89.39)	65.13 (69.45)	-2.44 (65.21)
SE Error (%)	20.87 (22.37)	-16.54 (15.98)	6.21 (18.13)
SOL Error (min.)	-15.36 (66.04)	12.14 (51.14)	-29.94 (59.99)

Abbreviations: Act., Actigraphy; SD, Standard Deviation.

Table 5- Sleep/wake identification performance in different disorder groups in the Test set.

		Total Accuracy Mean (SD)	Sleep Accuracy Mean (SD)	Wake Accuracy Mean (SD)	Kappa Mean (SD)	F1 Score Mean (SD)
Group 1	OA	0.78 (0.07)	0.88 (0.10)	0.52 (0.20)	0.42 (0.13)	0.58 (0.16)
	CPD	0.76 (0.07)	0.76 (0.13)	0.73 (0.23)	0.44 (0.20)	0.62 (0.20)
Group 2	OA	0.78 (0.09)	0.84 (0.13)	0.55 (0.12)	0.42 (0.10)	0.60 (0.10)
	CPD	0.75 (0.12)	0.75 (0.13)	0.63 (0.27)	0.37 (0.30)	0.58 (0.26)
Group 3	OA	0.73 (0.09)	0.84 (0.13) *	0.54 (0.25)	0.35 (0.20)	0.59 (0.16)
	CPD	0.69 (0.18)	0.64 (0.23)	0.78 (0.18) *	0.39 (0.25)	0.64 (0.20)
Group 4	OA	0.75 (0.17)	0.78 (0.16)	0.67 (0.17)	0.44 (0.33)	0.64 (0.18)
	CPD	0.70 (0.14)	0.62 (0.21)	0.87 (0.02)	0.42 (0.21)	0.65 (0.08)

* Wilcoxon signed-rank comparison of two methods, 5% significance level.

Table 6- Sleep study statistic estimation performance in different disorder groups in the Test set.

		SE Error Mean (SD)	WASO Error Mean (SD)	SOL Error Mean (SD)
Group 1	OA	-0.07 (18.75)	-8.56 (78.67)	40.11 (55.86)
	CPD	1.67 (10.89)	5.33 (38.42)	-11.94 (54.67)
Group 2	OA	0.71 (5.16)	-16.00 (37.62)	31.64 (43.37)
	CPD	-8.73 (14.38)	47.21 (72.94)	-44.07 (86.40)
Group 3	OA	-1.68 (16.93)	-2.14 (69.85)	20.18 (16.49)
	CPD	5.54 (19.42)	-1.04 (74.88)	-19.53 (50.13)
Group 4	OA	9.02 (2.67)	-49.75 (15.20)	25.75 (5.30)
	CPD	17.73 (13.65)	-59.50 (53.74)	-21 (31.11)

Figure 1

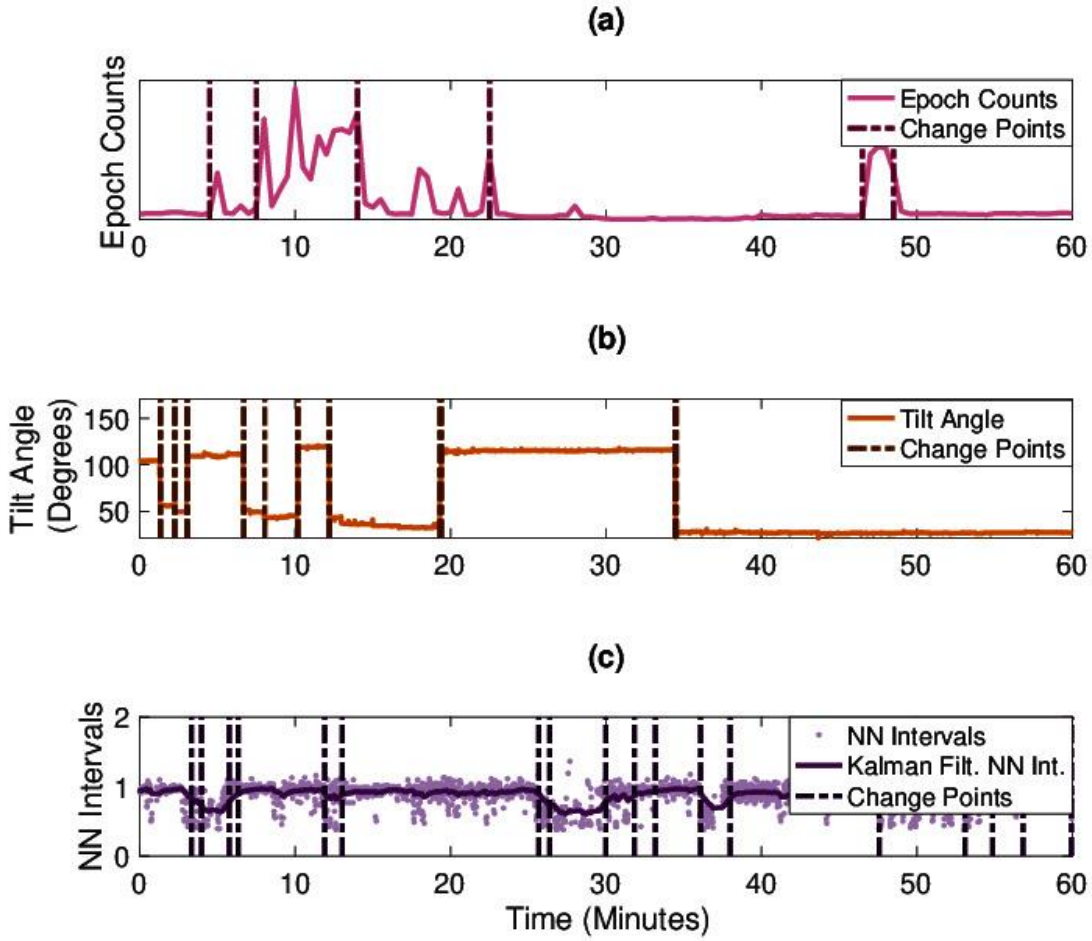

Figure 2

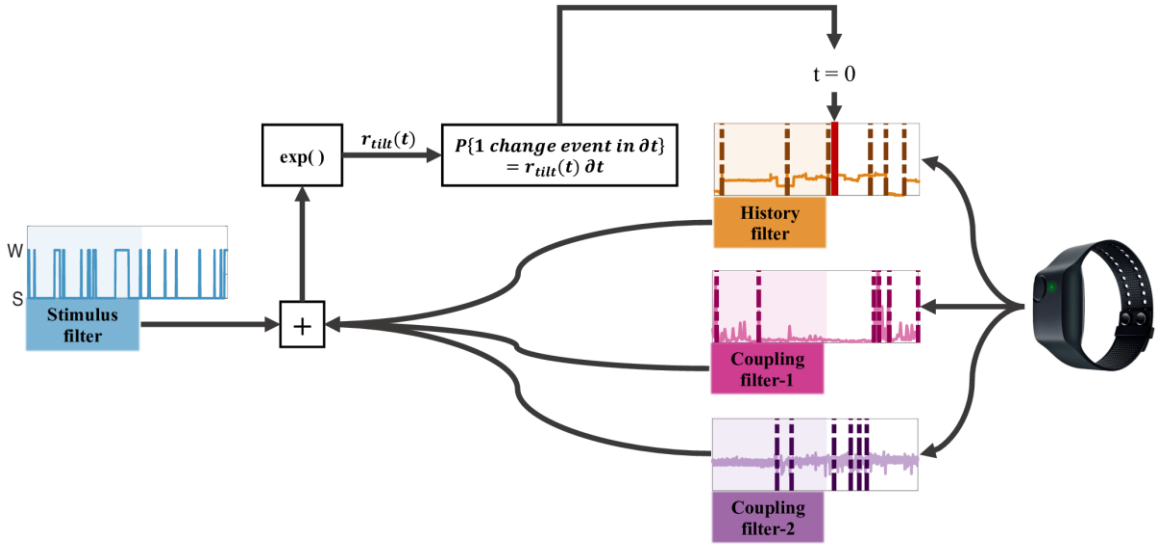

Figure 3

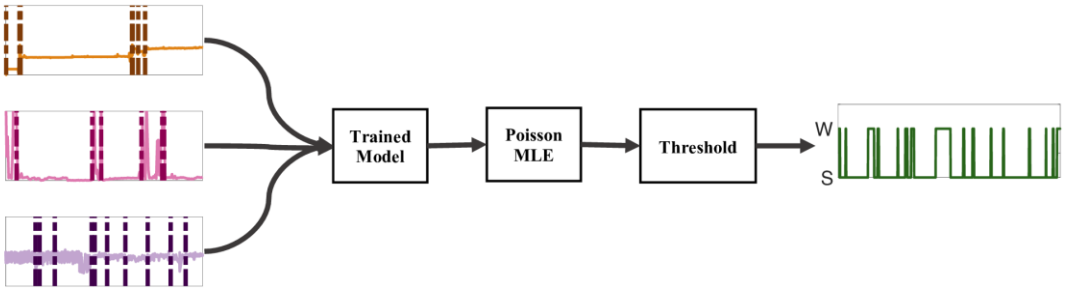

Figure 4

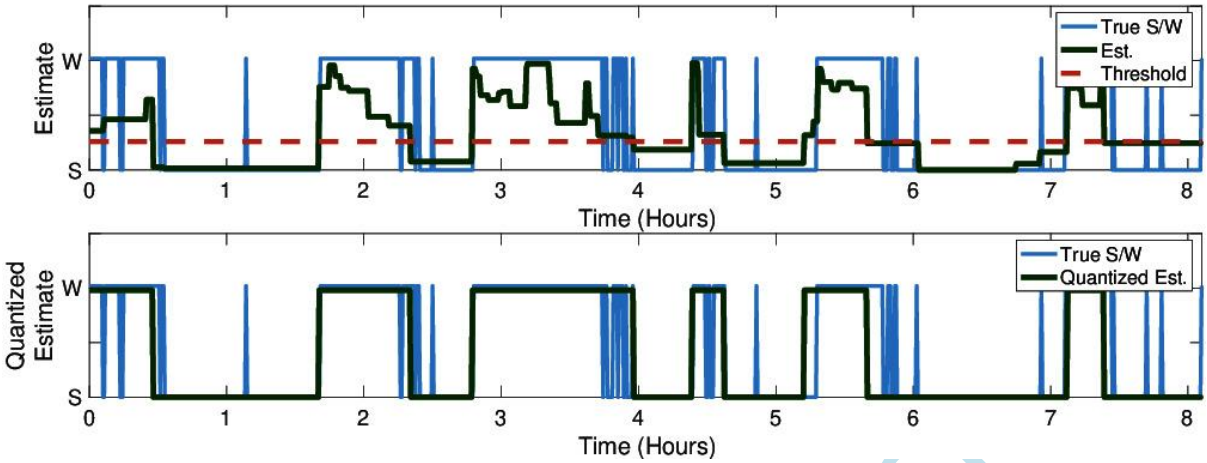

Figure 5

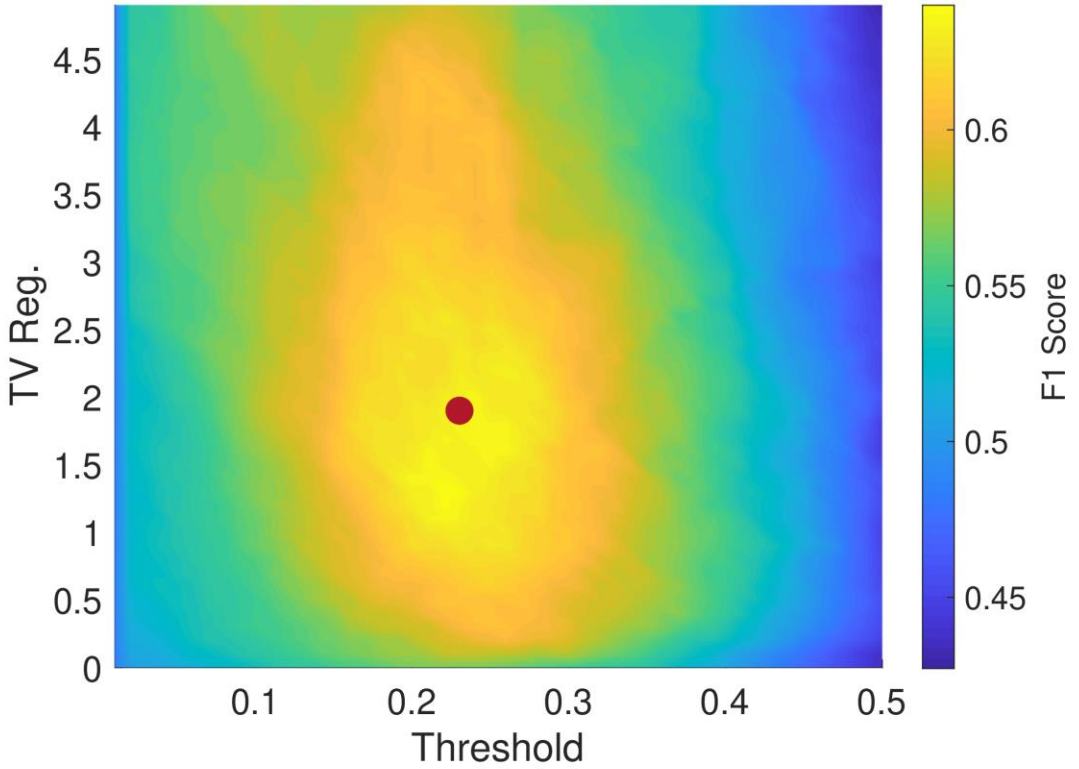

Figure 6

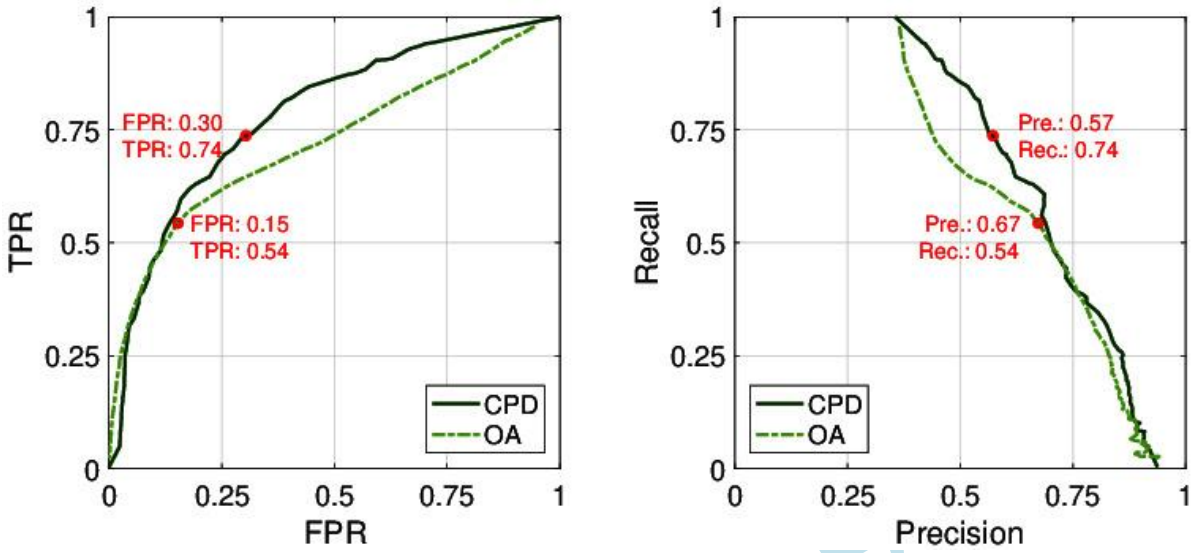

Figure 7

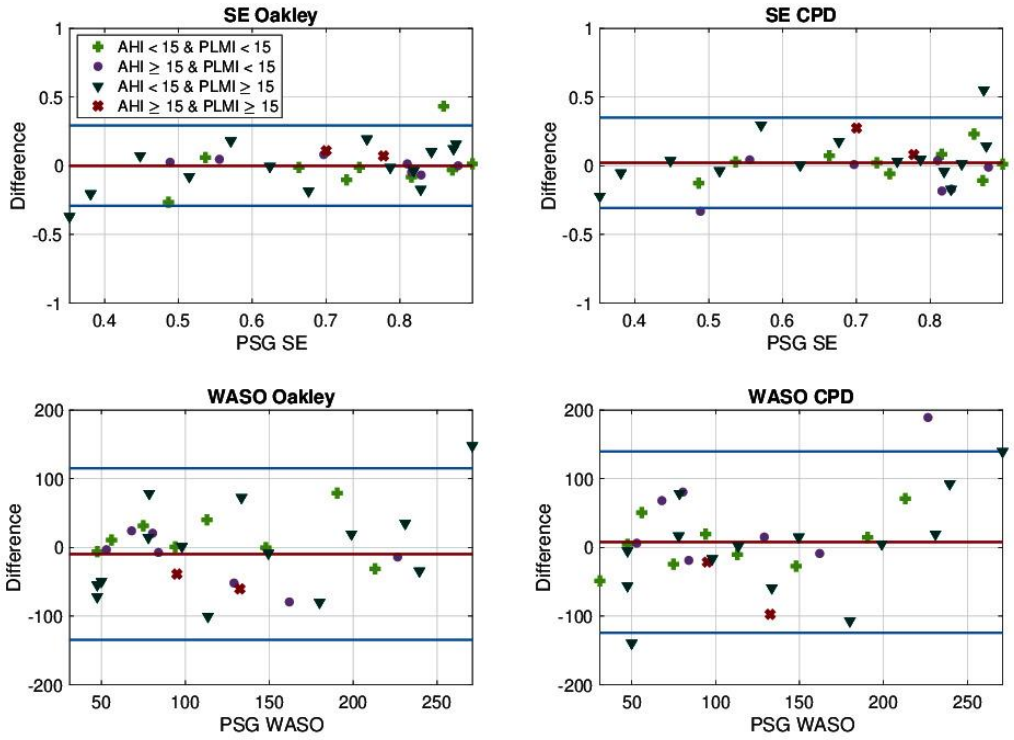

Figure 8