Detecting Sleep Using Heart Rate and Motion Data from Multisensor Consumer-Grade Wearables, Relative to Wrist Actigraphy and Polysomnography

Daniel M. Roberts[1], Margeaux M. Schade[2], Gina M. Mathew[2], Daniel Gartenberg[1], Orfeu M. Buxton[2-4]

[1]Sonic Sleep Coach, New York City, NY

[2]Biobehavioral Health, Pennsylvania State University, University Park, PA

[3]Division of Sleep Medicine, Harvard Medical School, Boston, MA

[4]Division of Sleep and Circadian Disorders, Departments of Medicine and Neurology, Brigham and Women's Hospital, Boston, MA

Corresponding author: Daniel M. Roberts; Mailing Address: Proactive Life Inc (DBA Sonic Sleep Coach) 175 Varick St, New York, NY 10014; Email: danmroberts@gmail.com

Abstract

Study Objectives: Multisensor wearable consumer devices allowing collection of multiple data sources, such as heart rate and motion, for the evaluation of sleep in the home environment, are increasingly ubiquitous. However, the validity of such devices for sleep assessment has not been directly compared to alternatives such as wrist actigraphy or PSG.

Methods: Eight participants each completed four nights in a sleep laboratory, equipped with polysomnography (PSG) and several wearable devices. RPSGT-scored PSG served as ground truth for sleep-wake state. Wearable devices providing sleep-wake classification data were compared to PSG at both an epoch-by-epoch and night level. Data from multisensor wearables (Apple Watch and Oura Ring) were compared to data available from ECG and a tri-axial wrist actigraph to evaluate quality and utility of heart rate and motion data. Machine learning methods were used to train and test sleep-wake classifiers, using data from consumer wearables. Quality of classifications derived from devices were compared.

Results: For epoch-by-epoch sleep-wake performance, research devices ranged in d' between 1.771 and 1.874, with sensitivity between 0.912 and 0.982, and specificity between 0.366 and 0.647. Data from multisensor wearables were strongly correlated at an epoch-by-epoch level with reference data sources. Classifiers developed from the multisensor wearable data ranged in d' between 1.827 and 2.347, with sensitivity between 0.883 and 0.977, and specificity between 0.407 and 0.821.

Conclusions: Data from multisensor consumer wearables is strongly correlated with reference devices at the epoch level and can be used to develop epoch-by-epoch models of sleep-wake rivaling existing research devices.

Keywords: Machine Learning, Big Data, Polysomnography, Actigraphy, Smartphone, Wearable, Artificial Intelligence

Statement of Significance:

Inexpensive and accessible multisensor wearable devices that allow for the collection of data relevant to sleep assessment, such as tri-axial accelerometer signals and heart rate, are increasingly common. Sleep-wake classifications from several wrist actigraph devices were compared against ground truth classifications derived from PSG. Machine learning was subsequently used to develop a novel classification algorithm of sleep-wake, using data from consumer multisensor wearables. The model that was developed using data from wearable devices generated more accurate epoch-by-epoch sleep-wake classifications than existing wrist actigraph devices, suggesting that such consumer devices may allow for the study of sleep among a more general population, and in more ecologically valid scenarios than was previously possible.

Introduction

The use of wrist-worn actigraphy, or watch-like devices sensitive to motion, to distinguish sleep from wake has been explored for over 40 years. The performance of a device to distinguish sleep from wake can be considered at two levels of analysis: the epoch-level, defined as the ability of a device to correctly classify each sleep epoch (typically 30 seconds) within the night, and the night-level, i.e. the ability of the device to summarize the entire night of sleep.

Kupfer and colleagues initially observed that at the night-level, movement counts derived from a wrist-worn actigraphy device were highly correlated with movement counts derived from polysomnography (PSG).[1] Wrist actigraphy has since emerged as a popular alternative to PSG because of the low obtrusiveness, more manageable cost, and adequate accuracy for monitoring sleep-wake states determined by PSG, the gold standard for monitoring sleep.[2] Despite actigraphy's long history and excellent ecological monitoring utility, however, devices that focus on motion-sensing suffer common limitations in wake detection using current algorithms.[3]

Human scorers of PSG or wrist actigraphy generate similar night-level summary statistics for sleep-wake, such as the total minutes of sleep within the night,[4] and can discriminate sleep from wake based on actigraphy data for as narrow as 1-minute sleep durations with high accuracy.[5] Automated staging algorithms using actigraphy data now classify sleep-wake, although with the continued need for a human scorer to define the "time in bed" period typically corroborated using sleep diaries.[6–8]

Common algorithms using wrist actigraphy data appear biased towards the classification of sleep,[3] with poorer accuracy for the detection of wakefulness. In terms of Signal Detection Theory treating sleep as the "target" to be detected within a sleep-wake classifier, actigraphy tends to be more accurate at correctly classifying periods of sleep (classifier sensitivity) than correctly classifying periods of wake (classifier specificity).[3] In practice, this detection imbalance means that actigraphy-based sleep-wake classifiers misclassify a greater number of wake epochs as sleep than vice-versa, leading to an over-classification of sleep. As a result, overall classification performance will be poorer for nights or individuals with a greater amount of wakefulness. For example, among individuals with sleep disorders, the bias against detecting wake is especially pronounced, with wakefulness detection of only 35%–50%.[3,9–14] As poor sleep impacts many aspects of daily life, including work productivity,[15–20] the risk of accidents,[21–25] chronic disease risk, and health care costs,[26–30] improving the assessment of sleep-wake outside of a laboratory environment is critical.

Commonly used actigraphy-based sleep-wake algorithms include the "Cole-Kripke"[6] and "Sadeh"[31] approaches, which use statistics on motion from both the epoch being classified, and from epochs several minutes into the past and future. As wrist actigraphy is conventionally implemented on a device dedicated to monitoring motion, some of the limitations in detecting periods of wake may be attributable to the use of a single monitoring modality. Many consumer wearable devices such as "smart watches" contain both tri-axial accelerometers and other sensor modalities. With their increasing ubiquity, there is an opportunity to noninvasively assess the synergy of multi-modal or multi-device-informed algorithm performance in sleep-wake classification.

Although sleep stages are typically assessed via changes in central nervous system (CNS), as measurable with electroencephalography (EEG), activity of the autonomic nervous system (ANS) also

varies by sleep stage and is measurable through changes in heart rate and heart rate variance over time. For example, de Zambotti et al.,[32] describes in a recent review the links between CNS and ANS activity with respect to sleep. The newest generation of wearable devices confers monitoring capabilities that make evaluation of ANS activity more accessible, primarily via the monitoring of cardiac activity via photoplethysmography (PPG) optical sensors. For the purpose of sleep-wake assessment, these PPG sensors provide a supplement to the accelerometers traditionally found in wrist actigraphs.

Some consumer wearables include the Oura Ring, Apple Watch, and Fitbit lines of devices. The Oura Ring, Apple Watch, and several models in the Fitbit line are capable of measuring both user motion and cardiac activity. In addition, some companies such as Oura and Fitbit have developed their own proprietary algorithms to generate measures of sleep quality from the raw data collected by their devices. Independent evaluations of these devices referenced against concurrently collected PSG have suggested accuracy for discriminating sleep-wake that is comparable to existing research-grade wrist actigraphy devices. For example, the Oura Ring has been reported to have a sensitivity to detect sleep of 96% and a specificity to detect wake of 48%,[33] whereas the Fitbit Charge 2 has been reported to have a sensitivity to detect sleep of 96% and a specificity to detect wake of 61%.[34] We refer to these new types of devices, which contain both accelerometers suitable for motion actigraphy and technology to measure cardiac activity (such as a PPG sensor), as *multisensor wearables*.

Beyond the algorithms provided by device manufacturers, many wearable devices also allow users or software developers to access the data from some or all of the sensors included on the device, via (for example) an application user interface (API) or software development kit (SDK). However, as noted by de Zambotti and colleagues[35], there is still room for improvement on the part of manufacturers. Specifically, device standardization, complete data access, and more algorithmic openness would benefit the research, clinical, and consumer communities. Allowing users and developers access to the raw sensor data permits the creation of novel algorithms and applications; for example, leveraging advances in machine learning to develop new classifiers beyond those provided by the manufacturer. In some cases, the sensor data from a device could be used by developers as it is collected to support 'real-time' detection and innovative interventions designed to enhance sleep.

Combining actigraphy with cardiac data has the potential to improve consumer-friendly devices beyond those which utilize motion data alone. Existing techniques such as the Cole-Kripke sleep-wake algorithm use motion alone, necessitating the need for development of new procedures to incorporate additional orthogonal data modalities. Supervised machine learning techniques are one method to flexibly incorporate new measurement modalities, such as the combination of motion and heart rate data, to distinguish sleep from wake. In the present study, several analyses were conducted to develop and evaluate the performance of a machine learning model relative to other common devices.

Within analysis 1, we first examined the quality of sleep-wake classification from several wearable devices, including common research wrist actigraphy devices and the Oura Ring, versus registered polysomnographic technologist (RPSGT) scored PSG as ground truth. Each device was evaluated in terms of how well it corresponded to PSG scoring at both an epoch-by-epoch and night-

level. The Apple Watch was not included within this first analysis, as it does not natively output sleep-wake classifications. Within analysis 2, we evaluated the quality of data available from two multisensor wearable devices, the Oura Ring and Apple Watch, by comparing their data to those obtained from an electrocardiography (ECG) channel of a PSG montage and to tri-axial motion data from a tri-axial wrist actigraph, the ActiGraph Link. This analysis was performed to establish the suitability of these devices for development of a sleep-wake classifier. Finally, within analysis 3, we used supervised machine learning techniques to develop and evaluate a series of sleep-wake classification algorithms informed by motion and cardiac data from the wearable devices. In order to evaluate how well similar models would compare using common laboratory data, models were also developed using data from the ECG channel of the PSG in conjunction with tri-axial motion data from a common research-grade wrist actigraph, serving as a point of comparison.

Methods

Participants

Healthy adults between 35 and 50 years of age with normal color vision and normal hearing were eligible to enroll. Participants were excluded for chronic or acute medical conditions requiring use of medication with reasonable likelihood to interfere with sleep or circadian structure, smoking within the past year, nocturnal shift work within the last 6 months or travel through more than 3 time zones within the last 3 months, substance use (verified with 9-panelmurine toxicology screening for amphetamines, cocaine, marijuana, opiates, phencyclidine, barbiturates, benzodiazepines, methadone, and propoxyphene at admission), and sleep disorders (screened during the first inpatient night). Participants were also asked to refrain from caffeine and alcohol for the three days immediately prior to inpatient admission. All participants provided written informed consent. All procedures were approved by the Institutional Review Board of the Pennsylvania State University and conducted in accordance with the Declaration of Helsinki.

Screening

Initial screening was conducted to evaluate typical participant sleep at home using a sleep diary and wrist actigraphy (Spectrum Plus, Philips-Respironics, Murrysville, PA) worn for at least four days and nights. Average sleep duration during the at-home screening was determined according to a previously described algorithm using 30-second epochs of movement count[3]; consecutive low-movement epochs were used to approximate sleep onset and awakening, and therefore sleep time each night. Each individual participant's average sleep duration from actigraphy was used as an alternate criterion for their minimum average sleep opportunity (time in bed) before later inpatient admission: participants received instruction to be in bed for 8 hours or their average, whichever was longer, for each of the 3 nights prior to lab admission.

Participants were confirmed to be free of sleep-related breathing disorders at home through pulse oximetry (Nonin Model 3150, Nonin Medical, Plymouth, MN); specifically, no nocturnal oxygenation fluctuations that resulted in substantial time spent below 88% peripheral oxygen saturation or 5 or more oxygen fluctuations per hour. Following sleep screening, participants underwent physical examination and medical history by a physician or nurse practitioner.

Protocol

Participants in the study completed four consecutive days and nights in the laboratory. On the 2nd and 4th night of participation, participants received auditory stimulation, designed to either enhance or disrupt sleep, referred to here as the Enhancing and Disruptive nights, respectively. The order of Enhancing and Disruptive stimulation was counterbalanced across participants. No auditory stimulation was presented on the 1st or 3rd nights, which are referred to as Habituation and Sham, respectively. The results of the auditory stimulation are outside the scope of this report, which is focused on the ability of various wearable devices to dissociate sleep from wake, and have been discussed elsewhere.[36] None of the devices or classification algorithms in this report had knowledge of experimental condition or the presence of auditory stimulation.

Participants were admitted to the Clinical Research Center at the Pennsylvania State University (University Park campus) and provided a biospecimen (urine) for toxicology screening. Participants were maintained in a light and sound attenuated sleep laboratory suite and instrumented with PSG (TrackIt v2.8.0.8, Lifelines Ltd., Irvine, CA; Polysmith v10.0 build 7956, Nihon-Kohden, Tokyo), comprised of EEG, electrooculography (EOG), electromyography (EMG), and ECG, all sampled at 200 Hz. The EEG montage contained 9 channels: ground, left mastoid (M1), right mastoid (M2), F3, Fz, F4, C3, Cz, C4, O1, and O2. Midline electrodes were referenced to the left mastoid, while lateral electrodes were referenced to the contralateral mastoid. ECG electrodes (two) were placed according to the AASM standard.[37]

In addition to PSG, participants were outfitted with several wearable devices, including an Apple Watch (wrist of non-dominant hand; distal), Oura Ring (best fitting finger per participant), ActiGraph Link (wrist of dominant hand), and Philips Respironics Spectrum Plus (wrist of non-dominant hand; proximal) (Table 1). Although the Apple Watch was placed on an alternate hand from the other two wrist devices, prior work has reported a high correspondence for wrist actigraphy concurrently collected from dominant and non-dominant wrists[38].

Participants were scheduled for a 9-hour time in bed sleep opportunity in darkness; due to technical and time constraints, actual sleep opportunity in the lab ranged from 8h 41m to 9h 13m (according to PSG, lights-out to lights-on).

[Insert Table 1 Here]

Polysomnography Processing and Sleep Staging

PSG data were staged in 30-second epochs according to American Academy of Sleep Medicine (AASM) standards by the RPSGT (author MMS). During staging, the RPSGT was blinded to the auxiliary data channel that included information regarding the timing presence of auditory stimulation. In some cases of discontinuous nocturnal recording due to unreadable EEG data (e.g., poor data quality or participant disconnected for restroom use), sections of data were considered "unscorable" and were excluded from epoch-level analyses.

Analysis 1: Evaluation of Device Classification

Sleep-wake classification from devices were compared to RPSGT-derived sleep staging to establish a point of reference for evaluating our machine learning model performance. Philips

Actiware software v6.0.9 was used to export the Actiwatch Spectrum Plus sleep-wake output at 30-second intervals. When classifying each 30-second sleep epoch, the Actiware software computes a weighted sum of activity counts within the current epoch, the prior 4 epochs, and the following 4 epochs. If this sum exceeds a pre-defined threshold, the epoch is classified as wake. Classifications were exported from the Actiware software using the default wake threshold of 'Medium,' corresponding to an activity count sum threshold of 40. ActiGraph ActiLife software v6.11.8 was used to export the ActiGraph Link sleep-wake output at 60-second intervals. The ActiLife software allows the selection of one of two algorithms, which are based upon (but may not be implemented identically to) existing sleep-wake algorithms; "Cole-Kripke" [6] or "Sadeh".[31] These two algorithm selections were separately output for comparison to our RPSGT staging. The Oura Ring (version 1) provides sleep staging with 4 categories (wake, "light" sleep, "deep" sleep, REM) at 30-second intervals. To compare the Oura output to other sleep-wake output, the 4-category staging was discretized into sleep and wake.

While the Actiwatch and ActiGraph devices score data continuously, the Oura Ring stores data when it determines the wearer is in bed. In some cases, the Oura Ring did not fully capture the period that the participant was in bed, and thus didn't begin to provide staging information by the time the experimental lights-out period began or stopped providing staging information before the experimental lights-out period ended. In these cases, those missing sleep epochs were filled to the bounds of the experimental in-bed period with the classification of "wake." In addition to the device classifications, a naïve "model" was constructed that always predicts sleep (because the "time in bed" period is typically skewed towards sleep rather than wake), serving as a point of reference for the performance of a theoretical classifier that provides no information.

The intervals used by each device for classification are not necessarily aligned to the intervals which define a 30 second sleep epoch labeled by the RPSGT. For example, even if the device provides classification in 30 second time windows, the start of the 30 second period may not be aligned to the start of the epoch used by the RPSGT for staging. For the purposes of comparison, each RPSGT-staged epoch was compared to the epoch from each device with the closest timestamp. In the case of data from the ActiGraph Link, while the device epochs are at 60-second intervals, the nearest labeled epoch to each 30-second RPSGT-staged reference epoch was used. If a device lacked a staged epoch within 30 seconds of a given PSG epoch due to missing data, that epoch was not evaluated for that device. Any epochs that couldn't be staged by the RPSGT (labeled "unscorable") were excluded from comparison. In addition, in order to mitigate the influence of clock offsets between each device and the PSG, which can especially influence the epoch-by-epoch comparisons, the sleep-wake output from each device was shifted +- 5 minutes relative to the PSG output, to identify the lag (if any) that optimized the sleep-wake correspondence between the device and PSG, for each night. This shifted version was subsequently used to compute both epoch-by-epoch and night-level statistics on data correspondence. A similar procedure (also with a +- 5 minute window) has been previously used to mitigate potential time offsets between actigraphy counts and sleep-wake staging[3].

Analysis 1A: Epoch-by-epoch correspondence between devices and PSG

Several metrics were computed using the Caret package for R[39] to evaluate the performance of the classifications from each device, treating the RPSGT-staged PSG as the correct or "ground truth" label for comparison, with sleep as the 'positive' class to be detected. These comparisons were performed after restricting to the complete cases of epochs that were present for both a given device, and for the PSG-derived staging; epochs that were missing for a given device or deemed unscorable by the RPSGT were not included in the comparison. The following metrics were evaluated:

Accuracy: the percentage of epochs correctly classified.

Balanced accuracy: the mean of wake epochs correctly classified and sleep epochs correctly classified.

Sensitivity: also known as recall, the percentage of sleep epochs correctly classified.

Specificity: the percentage of wake epochs correctly classified.

Precision: also known as positive predictive value, the percentage of epochs classified as sleep that are sleep.

Cohen's kappa (*κ)*: classifier agreement with PSG, relative to chance[40]. Specifically, kappa is calculated via $(p_o - p_e)/(1 - p_e)$, where p_o is the percentage of observed classifications with agreement, and p_e is the percentage of classifications that would be expected by chance.

Signal detection theory d-prime (d'): the difference in standard deviation between theoretical signal and noise distributions [41]. Computed by first transforming the percentage of sleep epochs correctly classified as sleep (classifier sensitivity, or 'hit rate') and the percentage of wake epochs incorrectly classified as sleep (classifier 'false alarm rate') into z-scores via the inverse of the normal cumulative density function, and then computing the difference.

Analysis 1B: Night-level correspondence between devices and PSG

For some applications, especially clinical, night-level summaries of sleep quality may be of greater interest than epoch-by-epoch sleep staging. For this reason, several metrics of sleep quality were also calculated, first using the ground truth values from the RPSGT-derived sleep staging, and then using the classifications provided by each device. These metrics were computed after restricting to the complete cases of epochs that were present for both the device being evaluated and the PSG-derived sleep staging:

Sleep efficiency: the percentage of the in-bed period classified as sleep.

Sleep onset latency (SOL): the latency, in minutes, from Lights Out until the first epoch of sleep occurs.

Total sleep time (TST): the total number of minutes of staged sleep.

Wakefulness after sleep onset (WASO): the number of minutes spent awake following the first epoch staged sleep.

For each device and classifier, these metrics were computed for all epochs present in the dataset and compared against the PSG staging data with the same epochs present.

Analysis 2: Evaluation of motion and heart rate data obtained from wearables

PSG Preprocessing

The ECG channel of the PSG was processed to serve as a point of comparison to the PPG data collected by the Apple Watch and Oura Ring, which each provide data that has already had some processing applied. The PSG data files were imported into MATLAB (Natick, MA), the ECG channel was selected, mean centered (removing any direct current offset), and bandpass filtered between 0.05 and 40 Hz, with a filter of order 20. The filtered time series was further decomposed using a symlet wavelet (sym4) to 5 levels, with the data at the 4^{th} and 5^{th} levels retained. In order to identify the R-peaks of the ECG QRS waveform, the square of the absolute value of the resulting wavelet-filtered time series was entered into a peak finding algorithm (the MATLAB function "findpeaks") with a minimum peak distance of 100 samples (.5 seconds) and a minimum peak height of 5 mV. Finally, the identified R-peaks were converted into R-R intervals by labeling each peak by the latency in seconds from the previous peak.

ActiGraph Link Preprocessing

Tri-axial accelerometer data from the ActiGraph Link, sampled at 80 Hz, was used as a point of comparison to actigraphy data derived from the Apple Watch and Oura Ring. In order to reduce the influence of gravity on accelerometer measurements, each dimension (x,y,z) in the accelerometer time series was high-pass filtered at 0.1 Hz with a 3^{rd} order Butterworth filter. Following filtering, each 3-element (x,y,z) sample was then converted to the magnitude of the 3-dimensional acceleration vector ($\sqrt{x^2 + y^2 + z^2}$).

Apple Watch Preprocessing

Although the Apple Watch contains a PPG sensor, access to the raw PPG signal was not available. Instead, an estimate of heart rate in beats-per minute (BPM) is provided by the device approximately every 5 seconds (at a rate of .2 Hz). The degree of processing performed by the device to transform the PPG signal to HR, or the time window used to compute each HR sample, is not documented by Apple, though it may be presumed that some temporal filtering is performed. As the Apple Watch provides values corresponding to heart rate, and not inter-beat interval (IBI), heart rate values were transformed to pseudo-IBI for consistency with the other devices, by dividing 60 seconds by the reported heart rate. Each Apple Watch sample was timestamped as it was collected from the device for synchronization with other devices in the study.

Tri-axial accelerometer data, corrected for the force of gravity, was additionally collected via the Apple Watch, sampled at 1.33 Hz (40 samples per 30 second sleep epoch). Although the device can be configured to sample at a higher rate, a higher sampling rate also requires more power, reducing battery life. A lower rate was used to ensure that the battery in the device would last through each night of data collection. The Apple Watch contains a tri-axial gyroscope, which allows the force of gravity to be separated from the raw accelerometer measurement, via knowledge of device orientation at each sample. The tri-axial, gravity-corrected time series was converted into a time-series of vector magnitude.

Oura Ring Preprocessing

The Oura Ring contains a PPG sensor (raw sensor signal unavailable). This signal is processed on the device to extract the R-peaks of the QRS waveform. When an R-peak is identified, the device logs the interval from the previous peak (the R-R interval) along with a timestamp.

The Oura Ring additionally contains a tri-axial accelerometer; however, the raw accelerometer data was not made available for the present study. Instead, a set of values summarizing the motion data were provided by the Oura device at 30-second intervals. Provided summary values included “motion seconds,” “motions low,” and “motions high.” Motion seconds is defined by Oura as the number of seconds in which an acceleration greater than 64 mg (where g is the acceleration due to gravity) occurs on any of the 3 accelerometer channels. Motions low and motions high are defined as counts, referencing the number of times the acceleration vector exceeds predefined threshold within the epoch; however, the actual thresholds exceeded are not provided by Oura. For each staged epoch, the summary value ‘motion seconds’ was used from the closest 30-second Oura provided epoch.

Temporal Alignment of Data

Data from each device were time-stamped during time of recording. In order to reduce the possibility of poor temporal alignment between devices, an additional alignment step was performed to align the cardiac activity data between the ECG, Apple Watch, and Oura Ring. For each recording and data source, outlier samples were removed, defined as an IBI outside the range of .4286 and 2 seconds (corresponding to an instantaneous heart rate range of 30 to 140 BPM), or more than 4 standard deviations from the mean inter-beat interval for a given recording and data source.

Comparing device data using raw samples is problematic, as a single missed IBI could shift all subsequent samples in the time series. In order to directly compare the timing of the devices, the IBI values derived from each device were linearly interpolated to a common sampling interval of integer seconds within the recording. To reduce the influence of outliers on the timing evaluation, an 11-sample median filter was applied to each interpolated time-series. Following the median filter, any chunk of data in which 10 or more contiguous seconds were missing (interpolated across) for either timeseries under comparison was removed from both timeseries. For the ECG and Apple Watch comparison, an average of 8.49 minutes of data were excluded per night, while for the ECG and Oura Ring comparison, an average of 18.93 minutes of data were excluded per night. Finally, the cross-correlation was performed between the ECG and device data at lags between 1000 and 1000 seconds to identify the timing offset that produced the highest correlation between the time series pair, separately for each recording.

A similar alignment process could not be performed for the actigraphy data, because unlike the ECG data, which is recorded as a channel within the PSG, no single source of actigraphy data could serve as ground truth with respect to PSG staging.

Data Comparison

The data from the Apple Watch and Oura Ring were compared to the data derived from the ECG or ActiGraph Link in order to evaluate how accurately the devices could capture variations in heart rate or motion across the night. To determine the quality of heart rate measurement, the mean IBI, SD of IBI, and RMSSD of IBI within each 30-second sleep epoch were correlated within each night between the ECG and Apple Watch, and separately between the ECG and Oura Ring. To determine the quality of motion actigraphy, the mean and standard deviation of actigraphy vector magnitude within each 30-second sleep epoch were correlated within each night between the ActiGraph Link and the Apple Watch. The magnitude of the three-element vector was used to compare accelerometer data instead of the three-coordinate time series because the vector magnitude is invariant to coordinate rotation between the devices under comparison. As raw motion actigraphy was not available for the Oura Ring (version 1) used here, the “motion seconds” variable instead correlated with the mean and standard deviation of vector magnitude as derived from the ActiGraph Link.

Analysis 3: Development and Evaluation of Machine Learning Approach

Feature Extraction

Each 30-second sleep epoch was labelled with the RPSGT sleep-wake determination as well as a set of features for the supervised machine learning model. The term ‘features’ refers to a set of computed values that are used as the input to the classifier. Features were extracted separately for each of three model variants constructed: 1) A model informed by PSG-derived ECG and actigraphy from the ActiGraph Link, 2) A model informed by cardiac activity and actigraphy from the Apple Watch, and 3) A model informed by cardiac activity and motion from the Oura Ring. In addition to the features derived from each device, a feature encoding the amount of time that has elapsed since the start of the “lights out” period was included, as the prior probability of a given epoch being sleep or wake was not assumed to be constant across the night.

Feature extraction began with the original data, not the interpolated data that was used to determine the cardiac temporal offsets. Temporal offsets for each recording (discussed earlier) were

applied to both the Oura Ring and Apple Watch data. Outlier rejection was again performed according to the guidelines mentioned previously.

For the combination of PSG and Link, and for the Apple Watch, the features within each epoch include the number of seconds that have elapsed since the "lights out" period began, the mean IBI in the epoch, the SD of IBI in the epoch, and the mean of actigraphy vector magnitude in the epoch. For the data derived from the Oura Ring, the variable "motion seconds" as output by the device were used in place of the mean of actigraphy vector magnitude. In addition to data from the current epoch, the mean IBI, SD of IBI, and mean actigraphy vector magnitude from the prior 8 epochs (4 minutes) were used as features within each current epoch, resulting in 27 cardiac or motion features and one time feature (the seconds elapsed since the in-bed period began) per epoch. The use of data from prior epochs when evaluating the current epoch was inspired in part by the Cole-Kripke actigraphy algorithm[6], which uses a weighted average of past and future epochs to determine the sleep-wake state of the current epoch. Here, only past and not future epochs are used in order to support the desired "real-time" prediction of sleep state.

Our machine learning approach used a gradient boosting classifier[42] as implemented within the Python package Scikit-learn.[43] Models were constructed and evaluated using nested leave-one-participant-out cross-validation. Within any machine learning paradigm, the data used to train the model should not be used to evaluate the model, motivating separate training and test partitions of the data.[44] In the case of grouped data, such as the present report where sleep epochs are grouped or clustered into nights and nights are grouped into participants, it is ideal if the test set contains more than a single or few participant instances, which may not be representative of the dataset as a whole. This leads to a leave-one-participant out cross-validation structure, in which each of the *n* participants serves as the test set within one of *n* model runs, and each model is trained on the remaining *n*-1 participants. The mean and standard deviation of performance across the *n* model runs is then collected, resulting in a value that is not biased towards any particular test participant.

However, many common machine-learning methods contain parameters that are not adjusted as part of the model, that instead govern the structure of the model itself. For example, for gradient boosted decision trees as used within the present report, such model structure parameters, or hyper-parameters, include the number of decision trees within the ensemble, the number of observations within the final terminating nodes, and others. A particular set of hyper-parameters may be optimal for a given dataset but are often unknown *a priori*. Iterating through model parameters to minimize the test set error generates a biased estimate of model performance, as test set performance is no longer outside of the training procedure.[45,46]

Within the nested leave-one-participant out cross-validation procedure employed here (Figure 1), selection of model parameters is implemented inside each outer cross-validation fold, via a separate series of (*n*-1) cross-validation folds per participant. Within each fold of the interior cross-validation, one of the (*n*-1) participants serves as a validation test for tuning the model hyper-parameters. For each outer cross-validation fold, the best candidate model from the interior cross-validation is then evaluated on the held-out test data. The final results reported are the mean and standard deviation of performance measures across the 8 outer cross-validation folds. In the present report, hyper-parameters and their options selected in the interior cross-validation include the learning rate (0.01, 0.1), the number of estimators (50, 100, 150), the maximum decision tree depth (1, 2, 3, 4), and the minimum number of samples per tree split (2, 10, 20, 30). Other model hyper-

parameters were fixed: loss, fixed at deviance, criterion, fixed at Friedman mean square error, and the minimum number of samples per leaf, fixed at 1. During training, models are optimized for AUC.

Models were separately trained and tested for each of the three datasets (PSG ECG channel in combination with ActiGraph Link accelerometer data, Oura Ring, and Apple Watch). For each dataset, gradient boosting classifier models were trained and tested with and without night-level normalization, and with and without class balance, totaling 4 variations per dataset. In all cases, the target of the classifier for each 30-second epoch was a binary classification of sleep or wake.

Normalizing data is a common preprocessing step within a machine learning paradigm, serving to place both features and datasets on a common scale. While scaling features is not required for a decision tree approach as described here, scaling between datasets can still provide a benefit when using physiological data such as heart rate, where dataset shift[47] may occur between training and test sets due to individual differences in mean heart rate. Here, night-level normalization refers to transforming each feature within each night by transforming to a z-score via subtracting the mean and dividing by the standard deviation of that feature within the night. However, normalizing at the night-level only allows classifications to be made once the full night is completed, as the distribution of values across the night is not known until the entire night of data is collected. Here, models were also run without normalization to simulate model performance when classifications are made in "real time," epoch-by-epoch, prior to the completion of the night.

"Class balance" refers to the balancing of sleep and wake instances used for model training (but not evaluation). Models were developed either with or without balanced classes during training. Class balance was achieved via random oversampling over the minority class (Wake) during training, via the Python package imbalanced-learn.[48]

The output of each classifier was evaluated in the same manner as the device output previously described, by restricting to complete cases between a given classifier and PSG-derived staging, at both the epoch-by-epoch and night-level. All of the epoch-by-epoch metrics previously described in the context of evaluating device output were calculated. In addition, classifier output probabilities were used to calculate the area under the curve (AUC) of the receiver-operating characteristic (ROC) curve. AUC could not be calculated for device output, such as from the Actiwatch Spectrum, etc. as these devices only provides discrete classifications, rather than class probabilities.

[Insert Figure 1 Here]

Results

Participants

Nine participants were enrolled in the protocol. One participant was excluded during the inpatient portion for medical attention not related to experimental interventions. The eight remaining participants completed the full, 4-night protocol (5 female, *M* = 40.75 years, *SD* = 4.84). Participants were adherent in maintaining their typical sleep schedule during the Pre-Inpatient

period (mean TST= 514.2m, *SD* = 59.8m), relative to Screening (*M* = 496.7m, *SD* = 44.4m), according to actigraphy, $t(7) = 0.95$, $p = .374$, $d = 0.33$.

Missing Data

Data from some of the 32 total nights of participation (8 participants with 4 nights each) were missing or corrupted for some datasets. Missing values were not imputed but were instead excluded from analyses. For data from the Apple Watch, three participants were each missing one night of data due to data corruption. In addition, the Apple Watch failed to accurately measure heart rate for 2 nights of a single participant, often oscillating between a value close to ECG-derived heart rate, and a value approximately twice the ECG-derived heart rate. These two nights were additionally removed from analysis. For Oura Ring data, two participants were each missing one night due to data corruption. In addition, two nights were missing large amounts of data, possibly due to an issue with the device's detection of the in-bed period, as described in the methods section. For one night, the device did not begin collecting data until approximately 2.5 hours following lights out. For a second night, the device was missing a contiguous chunk of approximately 2 hours of data within the night. These two nights were excluded from the comparison of night-level metrics for the Oura Ring but were included in the epoch-by-epoch sleep wake comparison, after excluding only the missing periods.

Epoch-by-epoch and night-level statistics were computed using the complete cases of epochs present for both a given device or classifier, and the PSG-derived sleep staging. For the Oura Ring staging output, no epochs were missing beyond the two datasets each missing 22-27% of data as described above. As previously described, while the Oura Ring sometimes began staging late, or ended staging early relative to the in-bed period, these periods were considered to have an Oura Ring classification of wake and thus weren't considered missing. No epochs were missing from the ActiGraph Link or Actiwatch Spectrum device output datasets. For the classification datasets, the Apple Watch datasets were missing a mean of 0.85% of epochs (SD = 2.15%), the ECG-Link datasets were missing a mean of 0.23% of epochs (SD = 0.34%), while the Oura Ring datasets were missing a mean of 4.01% of epochs (SD = 4.82%).

Data from one night of participation had a high percentage of epochs that were unscorable by the RPSGT (28%) due to a disconnected EEG electrode. This night was excluded from the comparison of night-level metrics. However, this night was included in epoch-by-epoch staging device comparisons and in our machine learning models, as these comparisons exclude any unscorable epochs from analysis. For the remaining 31 nights, the mean percentage of epochs marked unscorable by the RPSGT was 0.41% (*SD* = 0.64%).

Analysis 1: Evaluation of Device Classification

Analysis 1A: Epoch-by-epoch correspondence between devices and PSG

The sleep-wake classifications provided by several devices were compared to the values scored by the RPSGT (Table 2). For healthy individuals without a sleep disorder, the in-bed period is dominated by periods of sleep. In this case, overall classification accuracy (% of epochs correctly classified) can be a misleading performance metric. As illustration, Table 2 displays the epoch-by-epoch 'performance' that can be obtained by simply classifying every epoch within the in-bed period as Sleep, the "naïve" model. The sleep-wake classification accuracy of this no-information model is

near to each of the other devices used within the study. For metrics such as balanced accuracy or d-prime, the ActiGraph Link exported with the 'Sadeh' algorithm produced the highest concordance with RPSGT-derived staging. All classifiers perform better on the detection of sleep (sensitivity) than the detection of wake (specificity), as has been previously observed.[3] This bias was most pronounced for classifications from the Actiwatch Spectrum Plus, which produced both the greatest sensitivity for sleep epochs, and the poorest specificity for wake epochs. Despite this imbalance, as epochs within the in-bed period are predominantly sleep, the classifications from the Actiwatch Spectrum Plus produced the highest overall accuracy.

[Insert Table 2 Here]

Analysis 1B: Night-level correspondence between devices and PSG

The sleep-wake classifications provided by each device were used to compute several night-level metrics, including sleep efficiency, sleep onset latency, wake after sleep onset, and total sleep time, which were compared to the same metrics computed from the PSG-derived staging. Measures of device error are displayed within Table 3. Figure 2 displays a scatterplot of the correspondence between device and PSG-derived staging for each metric, while Figure 3 displays the bias of each device relative to PSG derived-staging for each metric. No device consistently outperforms the other across all metrics, and the best performing device in terms of RMSE for a given metric is not necessarily the same as the best performing device in terms of R^2. The Actiwatch Spectrum Plus results demonstrate the device's bias towards the classification of sleep. For example, Figure 2 illustrates that the Spectrum underpredicts SOL and WASO, while overpredicting TST and SE. The device biases within Figure 3 further demonstrate that this bias is not constant; for example, sleep efficiency is overpredicted to a greater extent for nights with lower sleep efficiency.

[Insert Table 3 Here]

[Insert Figure 2 Here]

[Insert Figure 3 Here]

Analysis 2: Evaluation of motion and heart rate data obtained from wearables

The Apple Watch failed to accurately measure the heart rate for two nights of a single participant. The Apple Watch data for those nights for that participant were excluded. Within these nights, the Apple Watch-derived heart rate fluctuated between a value that was either correct, or twice the rate of the participant's ECG derived heart rate. For the remaining recordings, the heart rate data from each device was aligned based on the identified offsets from the ECG time series. For the Oura Ring, the offset relative to the ECG ranged between -2 and 4 seconds (mean = 1.53, *SD* = 1.69) where negative values represent the device time leading the ECG, and positive values represent the device time lagging the ECG. For the Apple Watch, the offset relative to the ECG ranged between 0 and 12 seconds (mean = 8.43, *SD* = 3.31). The greater timing offset on the part of the Apple Watch may result from temporal smoothing of PPG data performed on the device, though the details of any preprocessing performed are not provided by Apple.

IBIs (or pseudo-IBIs in the case of the Apple Watch) that were outliers or extreme values were removed prior to feature extraction. For the ECG datasets, an average of 0.06% of samples

were rejected (*SD* = 0.19%). For the Oura Ring datasets, an average of 2.47% samples were rejected (*SD* = 2.53%). For the Apple Watch datasets, an average of 0.01% of samples were rejected (*SD* = 0.03%).

The epoch level feature values were compared between the ECG (for cardiac activity features) and ActiGraph Link (for motion feature) and the two wearable devices in order to determine the wearable device data quality. For the comparison of the Oura Ring with ActiGraph Link motion, the 'motion seconds' provided by the Oura Ring within a given epoch was compared to the mean and SD of vector magnitude from the ActiGraph Link, as raw accelerometer data was not available from the Oura Ring. For both the Apple Watch and Oura Ring, both heart rate and motion data from the wearables was reasonably correlated with data from the reference devices at the epoch level (Table 4). For example, the correlation of IBI with the ECG channel of the PSG montage averaged 0.92 and 0.85 for the Apple Watch and Oura Ring, respectively.
[Insert Table 4 Here]

Analysis 3: Development and Evaluation of Machine Learning Approach

Epoch-by-epoch sleep-wake classifications collected from the outer loop of the nested cross-validation procedure were compared to PSG derived staging, for each classifier variation under evaluation. Classifiers varied in their dataset, whether data was normalized within a night, and whether class oversampling was performed, resulting in 12 separate models. Table 5 displays the epoch-by-epoch performance of all model variations on several performance metrics. Data sets using heart rate data from the ECG channel of the PSG in combination with actigraphy data from the ActiGraph Link generally outperform classifiers using data from either the Apple Watch or Oura Ring. Normalization of data within a night generally improves classifier performance. Oversampling of wake epochs during training reduces much of the bias towards the classification of sleep epochs, which improves specificity to detect wake epochs while decreasing sensitivity to detect sleep epochs.

Figure 4 displays the mean ROC curve across nights, for the three datasets under study, both with and without night-level normalization. Model variations with class oversampling are not displayed, as AUC is largely unaffected by class oversampling.
[Insert Table 5 Here]
[Insert Figure 4 Here]

Night-Level Metrics Derived from Classifier Output

The sleep-wake output provided by each classifier were used to compute several night-level metrics, including sleep efficiency, sleep onset latency, wake after sleep onset, and total sleep time, which were compared to the same metrics computed from the PSG-derived staging. Measures of classifier error are displayed within Table 6. While the classifiers perform reasonably well at reproducing sleep efficiency, WASO, and total sleep time, they perform poorly at reproducing sleep onset latency. Figure 5 displays a scatterplot of the correspondence between classifier and PSG-derived staging for each metric, most evident is the difficulty of capturing the variance of sleep onset latency across nights. Figure 6 displays the bias of each classifier relative to PSG-derived staging for each metric.

[Insert Table 6 Here]

[Insert Figure 5 Here]

[Insert Figure 6 Here]

Discussion

Both epoch-by-epoch and night-level measures of sleep-wake detection were evaluated for two wrist devices and one ring device, relative to "gold standard" RPSGT-scored polysomnography. Consistent with past work that evaluated actigraphy for sleep-wake detection,[3] typical actigraphy and multisensor algorithms were of high sensitivity (% of sleep epochs correctly classified) but low specificity (% of sleep epochs correctly classified). Sleep, the more prevalent class within the in-bed period (approximately 88% of staged epochs in the present experiment), was over-classified, while wake was under-classified. This imbalance in performance is consistent with previous reports for the sleep-wake output of the Actiwatch Spectrum[3] and Oura Ring.[33]

The night-level analysis, comparing how accurately data from each device can be used to derive measures of sleep efficiency, sleep onset latency, WASO, and total sleep time, provides a complementary view of device performance. For example, although no single device was consistently optimal for all measures, the Oura Ring generated the highest R^2 relative to PSG-derived staging for 3 of the 4 measures under study, despite not necessarily generating the lowest mean error, and having lower performance than the ActiGraph Link in terms of the epoch-by-epoch performance measures of balanced accuracy or d'. While poor time synchronization between a given device and PSG could contribute to a discrepancy between epoch-by-epoch and night-level performance, a post-hoc synchronization process was performed within the present experiment to correct for any static time offsets between each device's sleep-wake output and PSG. Specifically, the sleep-wake output from each device was shifted +- 5 minutes relative to the PSG-derived staging, to identify the optimal correspondence.

Following analysis of the sleep-wake classifications output from the wrist actigraphy devices and Oura Ring, we evaluated the raw data quality from the Oura Ring and Apple Watch and developed our own novel machine learning models of sleep-wake. We also developed a classifier using data assumed to be of higher quality, using heart rate data derived from the ECG channel of the PSG in conjunction with tri-axial actigraphy data from the ActiGraph Link, for comparison.

In terms of raw data quality, we observed moderate to strong correlations between the data from the wearables and data from clinical measurement tools at the level of the 30-second sleep epoch. For example, despite the Apple Watch providing heart rate estimates approximately every 5 seconds, the mean pseudo-IBI within a sleep epoch computed from the Apple Watch had an average correlation of 0.92 to the mean IBI derived from the ECG channel of the PSG. Similarly, the mean IBI from the Oura Ring has an average correlation of 0.85 to the mean IBI from the ECG channel of the PSG.

Actigraphy features were also reasonably correlated despite some differences in the underlying data collection parameters. For example, the mean vector magnitude within an epoch derived from tri-axial accelerometer collected from the Apple Watch at 1.33 Hz had an average correlation of 0.83 to the mean vector magnitude collected from the ActiGraph Link at 80 Hz.

Similarly, the 'motion seconds' value provided by the Oura Ring at 30-second intervals had an average correlation of 0.53 to the mean vector magnitude from the ActiGraph Link. The lower correspondence for the motion feature derived from the Oura Ring does not necessarily suggest that the Oura Ring is poorer at capturing motion than the Apple Watch, but likely reflects the lower granularity of the data that was available for export from the Oura Ring ('motion seconds' at 30-second intervals) during the time of the study. In general, the high correspondence of data from these multisensor devices relative to PSG or high-resolution tri-axial actigraphy is promising with respect to using the devices when PSG is not available, and is applicable to the goal of longitudinal, non-invasive measurement of sleep.

We trained and tested variations of our classifier, using data from either the Apple Watch, Oura Ring, or for a point of comparison, data from the ECG channel of the PSG in combination with actigraphy data from the ActiGraph Link. In addition to the data features from each device, a temporal feature encoding the number of seconds that had elapsed into the in-bed period was included. Our temporal feature is linear, increasing in 30-second increments with each epoch. Other recent reports have also included temporal features in sleep staging models. For example, Fonesca et al.[49] used a linear time feature, while Walch et al.[50] used a mathematical model of the circadian clock specific to each participant. Each model was additionally trained with variations that incorporated class balancing via random oversampling of the minority class (wake), and/or normalization of datasets at the night-level via transformation of each feature to a z-score within the night. Oversampling was intended to ameliorate the common behavior of sleep-wake classifiers in the performance imbalance between the sensitivity and specificity.

Over sampling to balance classes during training did improve specificity (20-35%), but at a cost of reducing sensitivity (8-12%). Whether sensitivity or specificity are more favored may depend on application of the researcher. For example, because the oversampling procedure detects a greater percentage of wake epochs, it may be more appropriate to use in populations who have a sleep pathology. Practically, as oversampling did not affect AUC for the paradigm used here and our classifier natively outputs probabilities, a similar result could be achieved by shifting the probability threshold.

Night-level normalization resulted in a small improvement in model performance for all models with the exception of the model using Oura Ring data without oversampling. While normalizing data at the night-level can only be performed once the night has ended and the distribution of values across the night are known, using the distribution of data for a given user from prior session may allow for both data normalization and real-time classification.

The devices included in Analysis 1 vary along several dimensions, including the type of physiological data they use, their hardware implementation, and the algorithm internally used to generate sleep-wake classifications. However, despite these variations, they serve as a useful point of comparison for evaluating the performance of the alternative sleep-wake classifier we developed within Analysis 3.

The best performing research device for measuring epoch-by-epoch sleep-wake within the experiment as evaluated by d' was the ActiGraph Link running the Sadeh algorithm (d' = 1.874, sensitivity = .912, specificity = .647). For comparison, the classifier we developed, which was trained and tested using data from the Apple Watch with our normalization algorithm over the night

achieved epoch-by-epoch sleep-wake classification performance (d' = 2.347, sensitivity = 0.976, specificity = 0.602, AUC = .926). The same model with the addition of balancing classes prior to training results had a similar d', but a shift in bias away from sleep and towards wake (d' = 2.237, sensitivity = 0.898, specificity = 0.807, AUC = . 922).

Comparison of the performance of each device, and our models, on night-level summary measures such as sleep efficiency, sleep onset latency, WASO, and total sleep time revealed that the best performing device in terms of epoch-by-epoch sleep-wake classification is not necessarily the best for each summary measure. This is especially the case when considering that these measures can be evaluated in terms of either absolute accuracy (mean error), or their ability to capture variance across nights (R^2). For example, while the Oura Ring was not the most accurate device in terms of epoch-by-epoch accuracy among the research devices, it tended to capture the most variance relative to RPSGT-derived ground truth metrics (R^2), for all metrics except sleep onset latency. One caveat is that four nights were excluded from the Oura comparison; two nights were excluded due to corrupted recordings, and two were excluded due to a seeming failure of the device to detect the wearer was in bed and begin to provide classifications.

Similarly, while our models performed favorably in comparison to the research devices in terms of the epoch-by-epoch analysis, they were not more effective at deriving night-level summary values. As these summary values are not 'known' by the classifier during training, the optimal epoch-by-epoch classifier may not generate the optimal night summary, though in theory a perfect epoch-by-epoch classifier would reproduce all summary values precisely. One illustration of this discrepancy is that the classifier we developed using data from the Oura Ring often captured more variance between nights in terms of R^2 relative to the classifiers we developed using data from the Apple Watch or ECG-Link, despite having lower epoch-by-epoch performance. The data from our classifiers had particular trouble reproducing RPSGT staging-derived measures of sleep onset latency. As sleep onset latency according to AASM criteria is the time between 'lights out' and the first epoch scored as sleep, misclassification of a single early epoch can have a large influence on the accuracy of the measure.

A recent report by Walch et al.[50] also used data collected from an Apple Watch to develop classifiers for sleep-wake and sleep stage. Additionally, as previously described, this recent report also used a feature that encoded information about time within the night. However, there are several differences between this prior report and the present result. While Walch and colleagues collected tri-axial accelerometer data from the Apple Watch, accelerometer data was subsequently converted into movement counts using an algorithm developed to approximate Actiwatch movement counts from microelectromechanical systems (MEMS) accelerometers.[51] However, because the accelerometer contained within the Actiwatch is most sensitive to movements along a single axis (palmar-dorsal),[51] this approach utilizes only one axis of the original tri-axial accelerometer data, in support of approximating Actiwatch style movement counts. In contrast, our approach uses the vector magnitude of the tri-axial accelerometer data, the exception being the Oura Ring, for which we were only able to obtain count-level motion data from the device.

An additional distinction is that the approach described by Walch et al.[50] uses data from a 10-minute window around each sleep epoch being scored, while our approach uses only historical data, namely data from the 30-second epoch being scored in addition to data from the prior 8

epochs (4-minutes). Using data from the near future is common for sleep-wake staging; for example, the Cole-Kripke[6] and Sadeh[31] algorithms utilize movement counts from both the recent past and near future when classifying a given epoch, and is a reasonable approach if sleep-wake state is going to be evaluated retrospectively, after the night has completed. However, our approach instead facilities each 30-second epoch to be classified immediately, allowing for real-time applications.

Many multisensor wearable devices, such as the Apple Watch, allow data to be processed in real-time, as opposed to conventional actigraphy devices, which are designed to log the data on the device, and make the raw data accessible only after the sleep session has occurred. This precludes utility to inform real-time interventions that are delivered during monitoring with PSG. For example, ECG data collected in PSG was combined with pulse-oximetry to detect sleep stage and sleep apnea during sleep for the purpose of administering interventions to mitigate breathing obstruction.[52–54] The real-time availability of data in multisensor devices like the Apple Watch creates the possibility of delivering interventions during sleep for enhancing or manipulating sleep quality. Real-time sleep staging on accessible consumer wearables has several potential applications, including the treatment of sleep disorders,[52–54] enhancement of sleep with deep sleep stimulation,[55] and enhancement of memory through targeted memory reactivation.[56–58]

There are also notable limitations to the use of consumer wearable devices, in their present form. The battery life on a device such as the Apple Watch, while sufficient to collect data throughout a night of sleep, is currently not capable of 24 hours of continuous recording of high resolution data without needing to be recharged. Full access to raw sensor data from consumer wearables is often not available but could further improve the accuracy and transparency of sleep detection models. For example, the Apple Watch is equipped with a PPG heart rate sensor, but Apple does not currently allow developers to access the PPG sensor data. Instead, the PPG data is processed by the device and an estimate of instantaneous heart rate is provided at intervals of approximately 5-seconds. If the raw PPG sensor data was available from the Apple Watch, additional measures could be derived from the time series,[59] such as arterial blood pressure,[60] which could be used to improve the classification of sleep stage.[61]

There are several limitations to the present study. It included a relatively small sample of eight healthy participants, all in early mid-life, sleeping in a highly controlled laboratory environment. While it was important to include only healthy participants when investigating a small sample, additional work is required to identify how well sleep-wake models trained on healthy participants generalize to participants with sleep disorders and in outpatient settings. A potential solution is to add wearable devices to sleep studies already occurring, facilitating the relatively easy procurement of generalizing data to a broader cross section of individuals.

Additionally, although several data sources were included in the present report to demonstrate the relative performance that can be obtained with different sources of data, models were trained and tested within a given data source. More effort on data standardization may be required, for example, to train a model on PSG data and implement it using a wearable like the Apple Watch. Similarly, some devices provided sleep-wake classifications at 60-second intervals rather than the 30-second intervals in which the RPSGT staged the data, requiring alignment. Another limitation of the present study was that models were developed over the course of the night, as opposed to over a 24-hour period. A model developed on 24-hour data which contains a

greater number of periods of wakefulness may be useful to decrease the bias towards the classification of sleep that is common in sleep-wake classification. Further, while the present report focuses on classification of sleep-wake, a similar approach may be used to classify sleep stage.

The results of the study highlight the potential for using multisensor wearables to measure sleep in not only research and clinical contexts, but also in ecologically valid in-home study settings. Relative to many research-grade and clinical-grade actigraphy devices, consumer devices like the Oura Ring and Apple Watch are more affordable and accessible. Such devices are capable of sleep monitoring in the outpatient environment, which may be more reimbursable by healthcare companies than the expensive inpatient setting. The longitudinal, repeated, and multisensor quality of these devices make them ideal for evaluating insomnia or circadian rhythm disorders and can be further developed to provide users with tailored interventions that examine the cause of their sleep issues. Clinicians can then use these new tools to facilitate behavior change and administer more effective cognitive behavioral therapy for insomnia.[62]

Support:

Work was supported by the National Science Foundation (NSF) under grant #1622766 awarded to Gartenberg (PI; CEO Mobile Sleep Technologies LLC / Proactive Life Inc (DBA Sonic Sleep Coach). Work was conducted at Pennsylvania State University (via subcontract) and further supported by the Pennsylvania State University Clinical and Translational Sciences Institute (funded by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1TR002014) and institutional funds from the College of Health and Human Development of the Pennsylvania State University to Dr. Buxton. Collaboration also includes a separate project: NIH/NIA SBIR R43AG056250 to Gartenberg (PI; CEO Mobile Sleep Technologies LLC / Proactive Life Inc, DBA Sonic Sleep Coach) "Non-pharmacological improvement of sleep structure in midlife and older adults." Dr. Buxton also receives an honorarium from the National Sleep Foundation for his role as Editor in Chief (Designate) of Sleep Health (sleephealthjournal.org).

Acknowledgements:

We would like to thank the participants for providing their data.

Disclosures:

Financial Disclosures:

Daniel M. Roberts and Daniel Gartenberg are employed by Proactive Life, Inc. a for-profit company. Proactive Life has two patents granted or in application related to monitoring sleep:

Sleep Stimulation and Monitoring (US patent #14302482A).

Cyclical Behavior Modification (US Patent #8468115).

Orfeu M. Buxton discloses that he received two subcontract grants to Penn State from Mobile Sleep Technologies (NSF/STTR #1622766, NIH/NIA SBIR R43AG056250), received honoraria/travel support for lectures from Boston University, Boston College, Tufts School of Dental Medicine, and Allstate, and receives an honorarium for his role as the Editor in Chief (designate) of Sleep health sleephealthjournal.org.

Non-Financial Disclosures:

None

References

1. Kupfer DJ, Detre TP, Foster G, Tucker GJ, Delgado J. The application of delgado's telemetric mobility recorder for human studies. *Behav Biol*. 1972;7(4):585-590. doi:10.1016/S0091-6773(72)80220-7

2. Ancoli-Israel S, Cole R, Alessi C, Chambers M, Moorcroft W, Pollak CP. 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

3. Marino M, Li Y, Rueschman MN, 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

4. Kripke DF, Mullaney DJ, Messin S, Wyborney VG. Wrist actigraphic measures of sleep and rhythms. *Electroencephalogr Clin Neurophysiol*. 1978;44(5):674-676. doi:10.1016/0013-4694(78)90133-5

5. Mullaney DJ, Kripke DF, Messin S. Wrist-actigraphic estimation of sleep time. *Sleep*. 1980;3(1):83-92. doi:10.1093/sleep/3.1.83

6. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. *Sleep*. 1992;15(5):461-469. doi:10.1093/sleep/15.5.461

7. Jean-Louis G, Kripke DF, Cole RJ, Assmus JD, Langer RD. Sleep detection with an accelerometer actigraph: comparisons with polysomnography. *Physiol Behav*. 2001;72(1-2):21-28. doi:10.1016/S0031-9384(00)00355-3

8. Webster JB, Kripke DF, Messin S, Mullaney DJ, Wyborney G. An activity-based sleep monitor system for ambulatory use. *Sleep*. 1982;5(4):389-399. doi:10.1093/sleep/5.4.389

9. Blood ML, Sack RL, Percy DC, Pen JC. A comparison of sleep detection by wrist actigraphy, behavioral response, and polysomnography. *Sleep*. 1997;20(6):388-395.

10. de Souza L, Benedito-Silva AA, Pires MLN, Poyares D, Tufik S, Calil HM. Further validation of actigraphy for sleep studies. *Sleep*. 2003;26(1):81-85. doi:10.1093/sleep/26.1.81

11. Kushida CA, Chang A, Gadkary C, Guilleminault C, Carrillo O, Dement WC. Comparison of actigraphic, polysomnographic, and subjective assessment of sleep parameters in sleep-disordered patients. *Sleep Med*. 2001;2(5):389-396. doi:10.1016/S1389-9457(00)00098-8

12. 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

13. Signal TL, Gale J, Gander PH. Sleep measurement in flight crew: comparing actigraphic and subjective estimates to polysomnography. 2005;76(11):6.

14. Sivertsen B, Omvik S, Havik OE, 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

15. Chattu VK, Sakhamuri SM, Kumar R, Spence DW, BaHammam AS, Pandi-Perumal SR. Insufficient Sleep Syndrome: Is it time to classify it as a major noncommunicable disease? *Sleep Sci*. 2018;11(2). doi:10.5935/1984-0063.20180013

16. Hafner M, Stepanek M, Taylor J, Troxel W, Stolk C. *Why Sleep Matters -- the Economic Costs of Insufficient Sleep: A Cross-Country Comparative Analysis*. RAND Corporation; 2016. doi:10.7249/RR1791

17. Hillman D, Mitchell S, Streatfeild J, Burns C, Bruck D, Pezzullo L. The economic cost of inadequate sleep. *Sleep*. 2018;41(8). doi:10.1093/sleep/zsy083

18. Rosekind MR, Gregory KB. Insomnia risks and costs: health, safety, and quality of life. *Am J Manag Care*. 2010;16(8):11.

19. Rosekind MR, Gregory KB, Mallis MM, Brandt SL, Seal B, Lerner D. The cost of poor sleep: workplace productivity loss and associated costs: *J Occup Environ Med*. 2010;52(1):91-98. doi:10.1097/JOM.0b013e3181c78c30

20. Sluiter JK. High-demand jobs: Age-related diversity in work ability? *Appl Ergon*. 2006;37(4):429-440. doi:10.1016/j.apergo.2006.04.007

21. Bhattacharyya N. Abnormal sleep duration is associated with a higher risk of accidental injury. *Otolaryngol-Head Neck Surg*. 2015;153(6):962-965. doi:10.1177/0194599815604103

22. de Mello MT, Narciso FV, Tufik S, et al. Sleep disorders as a cause of motor vehicle collisions. *Int J Prev Med*. 2013;4(3):246-257.

23. Koppel S, Bohensky M, Langford J, Taranto D. Older drivers, crashes and injuries. *Traffic Inj Prev*. 2011;12(5):459-467. doi:10.1080/15389588.2011.580802

24. Pandi-Perumal SR, Verster JC, Kayumov L, et al. Sleep disorders, sleepiness and traffic safety: a public health menace. *Braz J Med Biol Res*. 2006;39(7):863-871. doi:10.1590/S0100-879X2006000700003

25. Tefft BC. Risks older drivers pose to themselves and to other road users. *J Safety Res*. 2008;39(6):577-582. doi:10.1016/j.jsr.2008.10.002

26. Léger D, Bayon V. Societal costs of insomnia. *Sleep Med Rev*. 2010;14(6):379-389. doi:10.1016/j.smrv.2010.01.003

27. Liu Y, Wheaton AG, Croft JB, Xu F, Cunningham TJ, Greenlund KJ. Relationship between sleep duration and self-reported health-related quality of life among US adults with or without major chronic diseases, 2014. *Sleep Health*. 2018;4(3):265-272. doi:10.1016/j.sleh.2018.02.002

28. Stenholm S, Head J, Kivimäki M, et al. Sleep duration and sleep disturbances as predictors of healthy and chronic disease–free life expectancy between ages 50 and 75: a pooled analysis of three cohorts. *J Gerontol Ser A*. February 2018. doi:10.1093/gerona/gly016

29. Wang S, Li B, Wu Y, et al. Relationship of sleep duration with sociodemographic characteristics, lifestyle, mental health, and chronic diseases in a large Chinese adult population. *J Clin Sleep Med*. 2017;13(03):377-384. doi:10.5664/jcsm.6484

30. Watson NF, Badr MS, Belenky G, et al. Recommended amount of sleep for a healthy adult: a joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society. *J Clin Sleep Med*. June 2015. doi:10.5664/jcsm.4758

31. Sadeh A, Sharkey M, Carskadon MA. Activity-based sleep-wake identification: an empirical test of methodological issues. *Sleep*. 1994;17(3):201-207. doi:10.1093/sleep/17.3.201

32. de Zambotti M, Trinder J, Silvani A, Colrain IM, Baker FC. Dynamic coupling between the central and autonomic nervous systems during sleep: A review. *Neurosci Biobehav Rev*. 2018;90:84-103. doi:10.1016/j.neubiorev.2018.03.027

33. de Zambotti M, Rosas L, Colrain IM, Baker FC. The sleep of the ring: comparison of the ŌURA sleep tracker against polysomnography. *Behav Sleep Med*. March 2017:1-15. doi:10.1080/15402002.2017.1300587

34. de Zambotti M de, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiol Int*. 2018;35(4):465-476. doi:10.1080/07420528.2017.1413578

35. de Zambotti M, Godino JG, Baker FC, Cheung J, Patrick K, Colrain IM. The boom in wearable technology: cause for alarm or just what is needed to better understand sleep? *Sleep*. 2016;39(9):1761-1762. doi:10.5665/sleep.6108

36. Schade MM, Roberts DM, Gartenberg D, Mathew GM, Buxton OM. Auditory stimulation during sleep transiently increases delta power and all-night proportion of NREM stage 3 sleep while preserving total sleep time and continuity. Presented at the: Society for Neuroscience; November 4, 2018; San Diego, CA. https://www.abstractsonline.com/pp8/#!/4649/presentation/38903. Accessed August 14, 2019.

37. Berry RB, Brooks, R, Gamaldo C, et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications*. Darien, IL: American Academy of Sleep Medicine; 2017.

38. Driller MW, O'Donnell S, Tavares F. What wrist should you wear your actigraphy device on? Analysis of dominant vs. non-dominant wrist actigraphy for measuring sleep in healthy adults. *Sleep Sci*. 2017;10(3):132-135. doi:10.5935/1984-0063.20170023

39. Wing MKC from J, Weston S, Williams A, et al. *Caret: Classification and Regression Training*.; 2019. https://CRAN.R-project.org/package=caret.

40. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas*. 1960;20(1):37-46. doi:10.1177/001316446002000104

41. Stanislaw H, Todorov N. Calculation of signal detection theory measures. *Behav Res Methods Instrum Comput*. 1999;31(1):137–149.

42. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232.

43. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res*. 2011;12:2825−2830.

44. Stone M. Cross-validatory choice and assessment of statistical predictions. *J R Stat Soc Ser B Methodol*. 1974;36(2):111-147.

45. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*. 2010;11:2079-2107.

46. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*. 2006;7(1):91. doi:10.1186/1471-2105-7-91

47. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. *Pattern Recognit*. 2012;45(1):521-530. doi:10.1016/j.patcog.2011.06.019

48. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res*. 2017;18(17):1-5.

49. Fonseca P, Teuling N den, Long X, Aarts RM. A comparison of probabilistic classifiers for sleep stage classification. *Physiol Meas*. 2018;39(5):055001. doi:10.1088/1361-6579/aabbc2

50. Walch O, Huang Y, Forger D, Goldstein C. Sleep stage prediction with raw acceleration and photoplethysmography heart rate data derived from a consumer wearable device. *Sleep*. doi:10.1093/sleep/zsz180

51. te Lindert BHW, Van Someren EJW. Sleep estimates using microelectromechanical systems (MEMS). *Sleep*. 2013;36(5):781-789. doi:10.5665/sleep.2648

52. Heneghan C, Chua C-P, Garvey JF, et al. A portable automated assessment tool for sleep apnea using a combined Holter-oximeter. *Sleep*. 2008;31(10):1432-1439.

53. Xie B, Hlaing Minn. Real-time sleep apnea detection by classifier combination. *IEEE Trans Inf Technol Biomed*. 2012;16(3):469-477. doi:10.1109/TITB.2012.2188299

54. Xie B, Qiu W, Minn H, Tamil L, Nourani M. An improved appoach for real-time detection of sleep apnea. In: *Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*. Rome, Italy: SciTePress - Science and and Technology Publications; 2011:169-175. doi:10.5220/0003137101690175

55. Ngo H-VV, Martinetz T, Born J, Mölle M. Auditory closed-loop stimulation of the sleep slow oscillation enhances memory. *Neuron*. 2013;78(3):545-553. doi:10.1016/j.neuron.2013.03.006

56. Oudiette D, Paller KA. Upgrading the sleeping brain with targeted memory reactivation. *Trends Cogn Sci*. 2013;17(3):142-149. doi:10.1016/j.tics.2013.01.006

57. Portas CM, Krakow K, Allen P, Josephs O, Armony JL, Frith CD. Auditory processing across the sleep-wake cycle: simultaneous EEG and fMRI monitoring in humans. *Neuron*. 2000;28(3):991-999. doi:10.1016/S0896-6273(00)00169-0

58. Rasch B, Buchel C, Gais S, Born J. Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science*. 2007;315(5817):1426-1429. doi:10.1126/science.1138581

59. Shelley KH. Photoplethysmography: beyond the calculation of arterial oxygen saturation and heart rate. *Anesth Analg*. 2007;105(On Line Suppl.):S31-S36. doi:10.1213/01.ane.0000269512.82836.c9

60. Teng XF, Zhang YT. Continuous and noninvasive estimation of arterial blood pressure using a photoplethysmographic approach. In: *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)*. Cancun, Mexico: IEEE; 2003:3153-3156. doi:10.1109/IEMBS.2003.1280811

61. Loredo JS, Nelesen R, Ancoli-Israel S, Dimsdale JE. Sleep quality and blood pressure dipping in normal adults. 2004;27(6):7.

62. Avidan AY. Sleep and neurologic problems in the elderly. *Sleep Med Clin*. 2006;1(2):273-292. doi:10.1016/j.jsmc.2006.04.010

63. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet Lond Engl*. 1986;1(8476):307-310.

Figure Captions

Figure 1. The nested cross-validation procedure used.

Figure 2. Comparison of PSG-derived and device-derived night-level sleep metrics. Points depict the PSG and device-derived values for each night of data within the study (except nights previously described as excluded). Lines depict the linear fit between PSG value and device value for each device.

Figure 3. Bland-Altman[63] plots comparing PSG-derived sleep metrics to the difference between each device-derived sleep metric and the PSG-derived sleep metric. Points depict difference values for each night of data within the study (except nights previously described as excluded).

Figure 4. Mean receiver operating characteristic (ROC) curves, for each classifier, both with and without night-level normalization. Each line depicts the point-by-point average of ROC across nights classified. Classifiers depicted are without class oversampling-

Figure 5. Comparison of PSG and classifier derived night-level sleep metrics. Points depict the PSG and classifier-derived values for each night of data within the study (except nights previously described as excluded). Lines depict the linear fit between PSG and classifier-derived values.

Figure 6. Bland-Altman[63] plots comparing PSG-derived sleep metrics to the difference between each classifier-derived sleep metric and the PSG-derived sleep metric. Points depict difference values for each night of data within the study (except nights previously described as excluded). Solid lines depict the mean bias between PSG and device-derived values, while dashed lines depict the 95% confidence interval for the bias across nights.

Tables and Table Captions

Table 1. Manufacturer, device name, and data details for wearable devices used within the study.

Manufacturer	Device	Data	Reference
ActiGraph (Pensacola, FL)	Link	Wrist worn device, tri-axial accelerometer sampling at 80 Hz.	http://actigraphcorp.com/products/actigraph-link/
Apple (Cupertino, CA)	Watch (Series 2)	Wrist worn device, tri-axial accelerometer sampling at 1.33 Hz, PPG sensor provides BPM estimates at approximately .2 Hz.	http://www.apple.com/watch/
Oura (Oulu, Finland)	Ring (Version 1)	Finger worn device, accelerometer internally sampling at 50 Hz, providing motion counts at 30 second intervals, PPG sensor internally sampling at 250 Hz, providing R-R intervals.	http://ouraring.com/
Philips Respironics (Murrysville, PA)	Spectrum Plus	Wrist worn device, accelerometer internally sampling at 32 Hz, providing motion counts at 30-second intervals.	http://www.actigraphy.com/devices/actiwatch/actiwatch-pro.html

Table 2. Comparison of device sleep-wake classification relative to ground truth sleep-wake classification derived from PSG staging. The "naïve" model is not a device but is instead the performance that is obtained by classifying every in-bed epoch as sleep, included here for reference. Values displayed within each cell are the mean and standard deviation across nights.

Source	Accuracy	Balanced Accuracy	D-Prime	Kappa	Precision	Sensitivity	Specificity
Naïve Model (result of always predicting Sleep, for reference)	0.876 (0.064)	0.500 (0.000)	0.646 (0.204)	0.000 (0.000)	0.876 (0.064)	1.000 (0.000)	0.000 (0.000)
ActiGraph Link with 'Cole-Kripke' algorithm	0.891 (0.046)	0.752 (0.070)	1.807 (0.375)	0.482 (0.120)	0.940 (0.042)	0.936 (0.050)	0.568 (0.163)
ActiGraph Link with 'Sadeh' algorithm	0.880 (0.054)	0.779 (0.071)	1.874 (0.463)	0.487 (0.146)	0.949 (0.041)	0.912 (0.064)	0.647 (0.163)
Actiwatch Spectrum Plus algorithm	0.904 (0.050)	0.674 (0.065)	1.831 (0.327)	0.424 (0.121)	0.914 (0.055)	0.982 (0.012)	0.366 (0.136)
Oura Ring	0.899 (0.046)	0.686 (0.075)	1.771 (0.502)	0.423 (0.150)	0.923 (0.042)	0.963 (0.039)	0.410 (0.165)

Table 3. Comparison of night-level metrics derived from device data, to those derived from RPSGT staging. Displayed are the mean error, mean absolute error (MAE), root mean squared error (RMSE), and R^2.

Metric	Device	Mean Error	MAE	RMSE	R^2
Sleep Efficiency (%)	ActiGraph Link (Cole-Kripke)	-0.989	4.038	5.472	0.396
	ActiGraph Link (Sadeh)	-4.032	5.248	7.580	0.348
	Actiwatch Spectrum	5.857	5.857	7.096	0.332
	Oura Ring	2.894	3.934	4.860	0.606
Sleep Onset Latency (min)	ActiGraph Link (Cole-Kripke)	-4.161	5.000	6.889	0.454
	ActiGraph Link (Sadeh)	-1.532	3.919	5.437	0.591
	Actiwatch Spectrum	-7.871	7.871	10.520	0.131
	Oura Ring	-1.462	5.308	8.863	0.297
WASO (min)	ActiGraph Link (Cole-Kripke)	9.613	20.613	29.245	0.473
	ActiGraph Link (Sadeh)	23.371	27.500	41.000	0.430
	Actiwatch Spectrum	-23.645	24.129	30.418	0.437
	Oura Ring	-14.096	20.750	23.671	0.670
Total Sleep Time (min)	ActiGraph Link (Cole-Kripke)	-5.452	21.839	29.676	0.386

	ActiGraph Link (Sadeh)	-21.839	28.387	41.105	0.340
	Actiwatch Spectrum	31.516	31.516	38.194	0.393
	Oura Ring	15.558	21.212	26.216	0.620

Table 4. Standardized correlation coefficients (*r*) for correlation of PSG ECG or ActiGraph Link-derived features, with features derived from the wearable devices under study. Values within each cell represent the mean and standard deviation (SD) across nights. IBI—inter-beat interval.

Device	Mean of Vector Magnitude (Apple Watch) or Motion Seconds (Oura Ring)	SD of Vector Magnitude (Apple Watch) or Motion Seconds (Oura Ring)	Mean of IBI	SD of IBI	RMSSD of IBI
Apple Watch	0.83 (0.08)	0.79 (0.08)	0.92 (0.04)	0.76 (0.09)	0.58 (0.18)
Oura Ring	0.53 (0.16)	0.47 (0.13)	0.85 (0.14)	0.78 (0.15)	0.62 (0.17)

Table 5. Performance of machine learning classifiers. Values in each cell are the mean and standard deviation across nights.

Dataset	Normalization	Oversampling	AUC	Accuracy	Balanced Accuracy	D-Prime	Kappa	Precision	Sensitivity	Specificity
Apple Watch	FALSE	None	0.918 (0.040)	0.918 (0.040)	0.753 (0.083)	2.228 (0.319)	0.544 (0.117)	0.936 (0.050)	0.974 (0.029)	0.532 (0.187)
	FALSE	Over Sampling	0.917 (0.037)	0.872 (0.077)	0.825 (0.060)	2.154 (0.377)	0.514 (0.145)	0.962 (0.038)	0.890 (0.095)	0.760 (0.156)
	TRUE	None	0.926 (0.040)	0.928 (0.029)	0.789 (0.063)	2.347 (0.373)	0.602 (0.102)	0.943 (0.039)	0.976 (0.022)	0.602 (0.136)
	TRUE	Over Sampling	0.922 (0.041)	0.882 (0.039)	0.853 (0.048)	2.237 (0.398)	0.533 (0.154)	0.967 (0.033)	0.898 (0.049)	0.807 (0.105)
Oura Ring	FALSE	None	0.897 (0.053)	0.914 (0.045)	0.692 (0.083)	1.926 (0.428)	0.441 (0.149)	0.929 (0.050)	0.977 (0.021)	0.407 (0.176)
	FALSE	Over Sampling	0.888 (0.063)	0.839 (0.075)	0.780 (0.079)	1.855 (0.451)	0.404 (0.135)	0.960 (0.039)	0.853 (0.096)	0.707 (0.208)
	TRUE	None	0.892 (0.063)	0.906 (0.050)	0.689 (0.082)	1.827 (0.454)	0.418 (0.135)	0.925 (0.051)	0.971 (0.031)	0.407 (0.174)
	TRUE	Over Sampling	0.896 (0.062)	0.844 (0.057)	0.805 (0.067)	1.876 (0.462)	0.422 (0.153)	0.964 (0.030)	0.855 (0.071)	0.755 (0.143)
ECG + Link	FALSE	None	0.924 (0.039)	0.917 (0.049)	0.755 (0.070)	2.321 (0.330)	0.560 (0.112)	0.938 (0.047)	0.970 (0.049)	0.540 (0.170)
	FALSE	Over Sampling	0.923 (0.037)	0.870 (0.082)	0.830 (0.055)	2.191 (0.350)	0.521 (0.140)	0.966 (0.035)	0.883 (0.105)	0.776 (0.144)
	TRUE	None	0.926 (0.041)	0.922 (0.043)	0.788 (0.065)	2.383 (0.436)	0.595 (0.113)	0.943 (0.040)	0.968 (0.045)	0.607 (0.147)
	TRUE	Over Sampling	0.924 (0.042)	0.878 (0.048)	0.856 (0.043)	2.253 (0.372)	0.536 (0.156)	0.970 (0.029)	0.890 (0.059)	0.821 (0.096)

Table 6. Estimation of night-level metrics, using data from classifiers built on various data sources. Here, the values reported are for the models computed with night-level normalization, but without class oversampling.

Metric	Device	Mean Error	MAE	RMSE	R^2
Sleep Efficiency (%)	Apple Watch	2.551	4.239	4.715	0.334
	Oura Ring	3.584	4.552	5.436	0.43
	ECG + Link	1.836	4.635	5.478	0.266
Sleep Onset Latency (min)	Apple Watch	-2.154	5.885	8.700	0.000
	Oura Ring	-1.379	3.793	5.019	0.039
	ECG + Link	-0.871	6.516	9.076	0.003
WASO (min)	Apple Watch	-11.577	18.269	20.485	0.506
	Oura Ring	-16.897	21.517	25.418	0.476
	ECG + Link	-8.952	19.113	23.794	0.421
Total Sleep Time (min)	Apple Watch	13.731	22.692	25.291	0.346
	Oura Ring	18.276	23.276	27.638	0.756
	ECG + Link	9.823	24.919	29.484	0.307

Figure_1

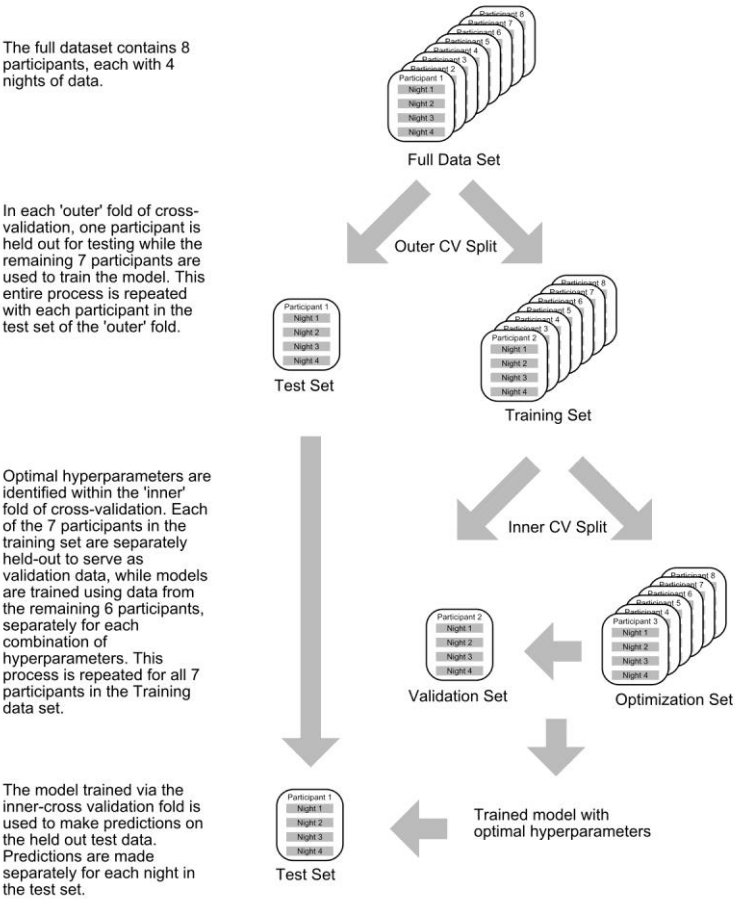

Figure_2

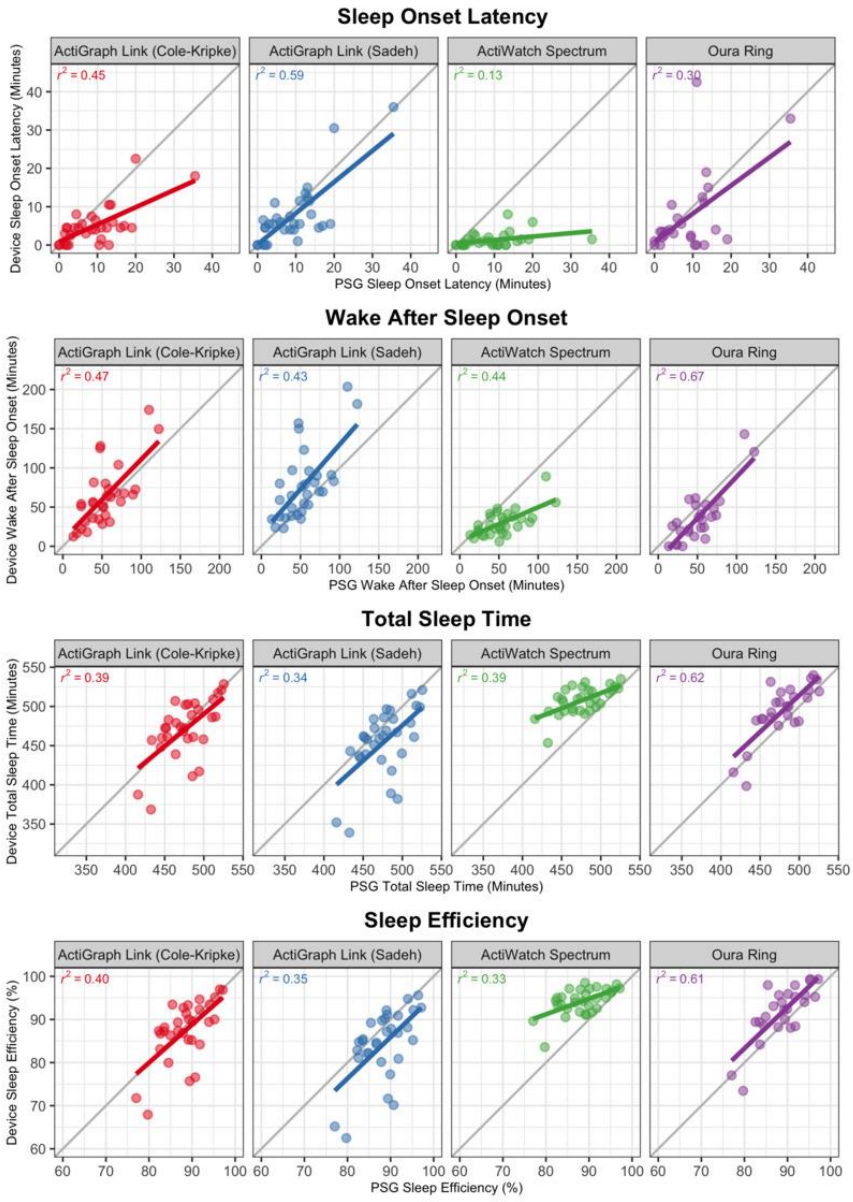

Figure_3

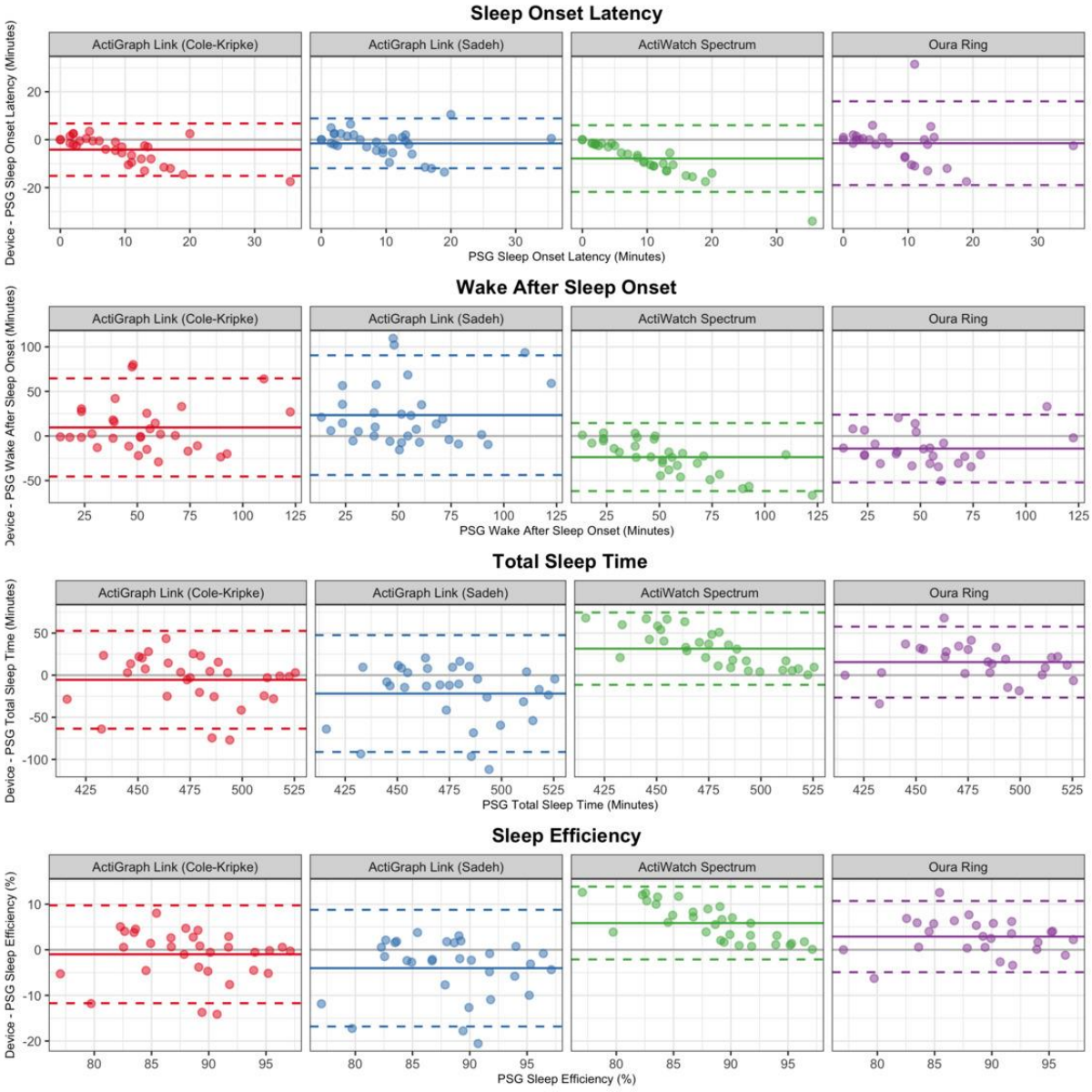

Figure_4

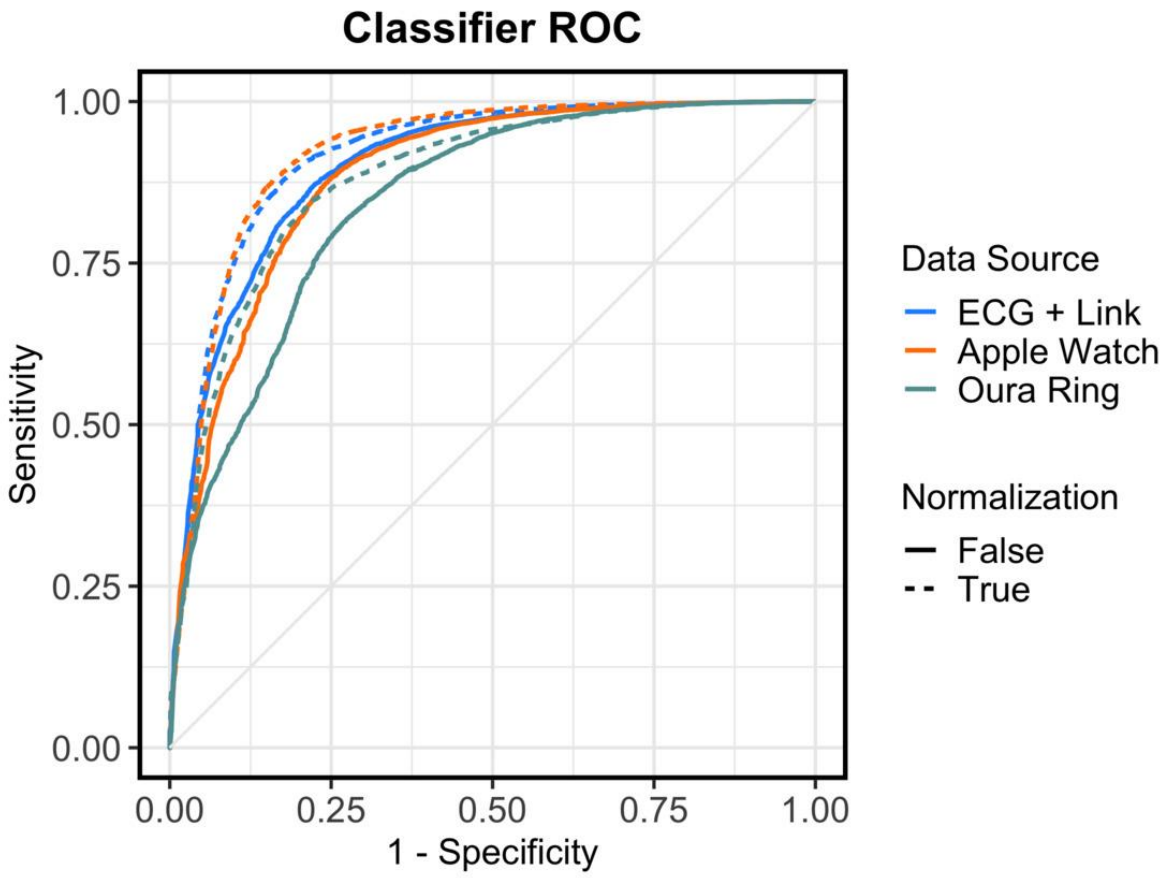

Figure_5

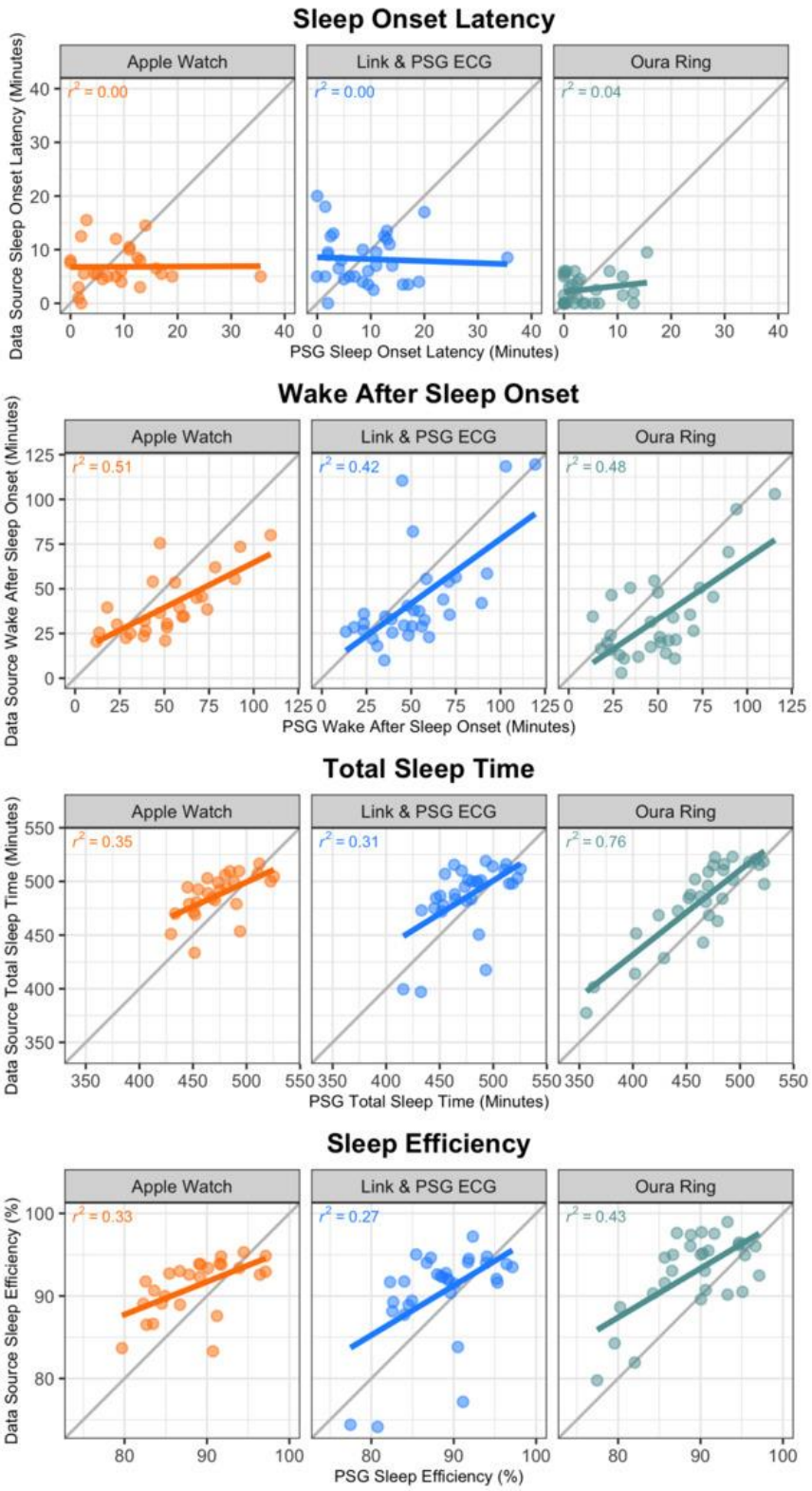

Figure_6