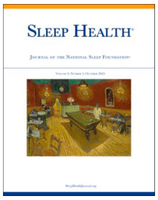
Contents lists available at ScienceDirect

Sleep Health: Journal of the National Sleep Foundation

journal homepage: www.sleephealthjournal.org



Performance of an open machine learning model to classify sleep/wake from actigraphy across ∼24-hour intervals without knowledge of rest timing

Daniel M. Roberts, PhD [a,b,*], Margeaux M. Schade, PhD [a], Lindsay Master, MAS [a], Vasant G. Honavar, PhD [c], Nicole G. Nahmod, MPH, MMS, PA-C [a], Anne-Marie Chang, PhD [a], Daniel Gartenberg, PhD [b], Orfeu M. Buxton, PhD [a]

[a] *Department of Biobehavioral Health, The Pennsylvania State University, University Park, Pennsylvania, USA*
[b] *Proactive Life, Inc, New York, New York, USA*
[c] *Faculty of Data Sciences, College of Information Science and Technology, The Pennsylvania State University, University Park, Pennsylvania, USA*

ARTICLE INFO

Article history:
Received 10 November 2022
Received in revised form 5 June 2023
Accepted 2 July 2023

Keywords:
Actigraphy
Machine learning
Classification
Wearables
24-hour

ABSTRACT

Goal and aims: Commonly used actigraphy algorithms are designed to operate within a known in-bed interval. However, in free-living scenarios this interval is often unknown. We trained and evaluated a sleep/wake classifier that operates on actigraphy over ∼24-hour intervals, without knowledge of in-bed timing.
Focus technology: Actigraphy counts from ActiWatch Spectrum devices.
Reference technology: Sleep staging derived from polysomnography, supplemented by observation of wakefulness outside of the staged interval. Classifications from the Oakley actigraphy algorithm were additionally used as performance reference.
Sample: Adults, sleeping in either a home or laboratory environment.
Design: Machine learning was used to train and evaluate a sleep/wake classifier in a supervised learning paradigm. The classifier is a temporal convolutional network, a form of deep neural network.
Core analytics: Performance was evaluated across ∼24 hours, and additionally restricted to only in-bed intervals, both in terms of epoch-by-epoch performance, and the discrepancy of summary statistics within the intervals.
Additional analytics and exploratory analyses: Performance of the trained model applied to the Multi-Ethnic Study of Atherosclerosis dataset.
Core outcomes: Over ∼24 hours, the temporal convolutional network classifier produced the same or better performance as the Oakley classifier on all measures tested. When restricting analysis to the in-bed interval, the temporal convolutional network remained favorable on several metrics.
Important supplemental outcomes: Performance decreased on the Multi-Ethnic Study of Atherosclerosis dataset, especially when restricting analysis to the in-bed interval.
Core conclusion: A classifier using data labeled over ∼24-hour intervals allows for the continuous classification of sleep/wake without knowledge of in-bed intervals. Further development should focus on improving generalization performance.

Introduction

Rationale

Clinical polysomnography (PSG) remains the gold (or "reference") standard for quantifying sleep. However, use of PSG is constrained by equipment cost, sleep technologist labor, and intrusiveness. Particularly for long-term ambulatory monitoring, it can be impractical for participants to be equipped with PSG. When

[*] Corresponding author: Daniel M. Roberts, PhD, Department of Biobehavioral Health, The Pennsylvania State University, University Park, PA, USA.
E-mail address: danmroberts@gmail.com (D.M. Roberts).

https://doi.org/10.1016/j.sleh.2023.07.001

dissociating intervals of sleep from intervals of wake is of primary interest, wrist-worn actigraphy devices are commonly used instead of PSG.[1]

Common actigraphy algorithms classify an epoch of time as either sleep or wake based on the magnitude of the activity count in the current and nearby epochs, with the Coke-Kripke,[2] Oakley,[3] Sadeh,[4] and Scripps Clinic[5] algorithms being among the most prominent. These algorithms have been developed with a focus on operating within a known or assumed rest or “in-bed” interval, and performance analyses of actigraphy classifications are also most often restricted to in-bed intervals. Within in-bed intervals these algorithms exhibit high sensitivity (classification accuracy on sleep epochs), but relatively low specificity (classification accuracy on wake epochs).[6,7] As sleep tends to predominate within the in-bed interval, this classification imbalance can generate high epoch-by-epoch accuracy, but also often leads to underestimation of wake time, especially for individuals with low sleep efficiency (SE).[8]

For many paradigms, the true in-bed interval is known, such as sleep laboratory studies. However, for many other situations, the true in-bed interval is not known a priori, such as when studying participants in their home. In these cases, it is common for researchers to first identify the candidate in-bed interval via some combination of sleep diary, participant-initiated event markers, visual inspection of actigraphy counts, or a rest-interval algorithm derived from activity counts. However, there are limitations to each of these methods.

Significance

To avoid the dependence on the identification of in-bed intervals and the associated bias toward the classification of sleep, we have trained and evaluated a sleep/wake classifier on data labeled over approximately 24-hour intervals.

Background

Most existing sleep/wake classifiers operate best within predefined in-bed intervals, often identified via participant self-report through sleep diary or event marker button press. Sleep diaries such as the Consensus Sleep Diary[9] have been reported to perform well at capturing the timing of the rest interval relative to single-channel EEG, with average bias of 6 minutes and 95% limits of agreement of 1.4 hours.[10] However, requiring a sleep diary or participant-initiated event marker for actigraphy analysis can lead to missing data if these measures are not consistently completed. Lauderdale et al[11] report that when participants were asked to track their sleep over the course of three nights, 40% of bedtimes or waketimes were not recorded by participants with the actigraphy event marker and needed to be imputed based on participant sleep diaries.

A further complication regarding the use of in-bed intervals defined by sleep diary is that many sleep diaries do not include enough detail to capture the timing of nonprimary rest intervals (ie, napping). For example, the Consensus Sleep Diary asks respondents two questions about napping, "How many times did you nap or doze?" and "In total, how long did you nap or doze?." However, as these questions do not ask respondents to indicate *when* they napped, they cannot be used to determine the timing of the nap required to apply an “in-bed” actigraphy algorithm. Participant-initiated event markers can likewise have issues capturing the onset of unintentional naps, which would need to be logged by participants retroactively and may not be recalled correctly. Correctly capturing unintentional naps is especially important for studying individuals with excessive daytime sleepiness[12] or populations of older adults, for whom naps are more common.[13]

Another consideration is maintaining independence between a subjective sleep diary and objective actigraphy data, as the discrepancy between subjective and objective sleep is clinically useful. For individuals with sleep disorders, some measures captured by actigraphy are unique from the same measures captured via sleep diary,[14] with individuals with insomnia tending to underestimate their sleep quality.[15-18] Subjective-objective sleep discrepancy is additionally greater and more variable night-to-night for older adults receiving cognitive behavioral therapy for insomnia relative to controls pretreatment, with the change in sleep discrepancy post-treatment correlated with treatment efficacy as assessed with Insomnia Severity Index.[19] Classifying sleep/wake at a ~24-hour interval without using in-bed timing derived from sleep diary maintains the independence of the two measures.

When self-report measures are not available, an alternative is to estimate the in-bed interval from the timeseries of activity counts, either through visual inspection or an algorithm designed to determine the time limits of the interval. Visual inspection is problematic due to limitations on the reproducibility of subjective assessments. Algorithms often contain assumptions about the length or number of in-bed intervals within in a day. For example, Kanady and colleagues report[20] the performance of an automatic minor rest interval algorithm (AMRI) for detecting the presence of daytime naps, across a combination of settings related to motion sensitivity and interval length threshold. Daytime naps could be detected via the automatic minor rest interval algorithm, with higher sensitivity settings producing the best correspondence with PSG. However, these high-sensitivity settings also tended to misclassify naps in recordings that contained no naps. The interval setting could be adjusted based on likelihood that an individual napped, but this is often unknown a priori.

Aims

An epoch-by-epoch sleep/wake classifier was trained and evaluated on data from multiple studies with ~24-hour PSG staging or direct participant observation of wakefulness, alongside a common wrist actigraphy device. Here ~24-hour refers to collecting data across the majority of a day, rather than surrounding a known in-bed period, although 1 of 4 studies recorded approximately 42 hours of data across 2 days, and the remaining studies don’t include precisely 24 hour of data due to periods in which the actigraphy device is off-wrist. The classifier, a temporal convolutional network (TCN) deep-learning approach, uses the time series of actigraphy activity counts across the day as input. The performance of the resulting classifier is evaluated using standard criteria[21,22] at the level of the 30-second sleep epoch, and at the level of the interval (eg, total sleep time [TST] across the interval). For comparison, the same metrics were compared against the Oakley classification algorithm native to the actigraphy device.

Although the primary goal was to develop a classifier for ~24-hour data without knowledge of in-bed intervals, the performance of the ~24-hour classifier was additionally evaluated within the limits of the known primary in-bed intervals. This was done to establish how the classifier, which is trained on data composed primarily of wake, would perform against alternatives in an interval composed primarily of sleep, such as when the in-bed interval is identified via alternate methods. For example, if the ~24-hour classifier also performs adequately within a known in-bed interval, a single classification approach could be used in both contexts. Additionally, the model was evaluated on the Multi-Ethnic Study of Atherosclerosis (MESA) dataset, to estimate performance on a different paradigm than used to train the model.

Methods

Sample

Data from four studies where sleep/wake was staged or verified as wake over an approximately 24-hour interval were incorporated

Table 1
Description of data collection for each data set used in analysis

Dataset name	ActiWatch model	ActiWare software version	Data collection setting	Observation interval	PSG interval	Day interval	Used for	Description
Deep Sleeping	Spectrum Plus	6.0.9	Laboratory	24-h	In-Bed	4-4 PM	Training, Tuning, Testing	Laboratory study of acoustic stimulation during sleep. Data from 12 participants over 3 days and nights; each session of data collection separated by ≥2 nights. Protocol has not been previously reported.
EcoSleep	Spectrum Plus	6.0.9	Home	None	~42-h	Variable, using available data split into 2 equal-sized intervals for 18 of 19 participants, and 1 31-h interval for the remaining participant due to partial data loss, see Description column. 8:15 PM-5:47 PM on average	Training, Tuning, Testing	In-home portable PSG study with data from 19 participants. Data recorded over 42 h on average, including 2 nocturnal periods for 18 of 19 participants. Due to data loss, one participant was monitored for a period of 31 h, with only 1 nocturnal period. Protocol has not been previously reported.
MESA	Spectrum (Classic)	5.59	Home	None	Primarily in-bed, with limited data preceding and following	Variable, using all staged data. 9:31 PM-6:24 AM on average	Testing	In-home study with 1 night of unattended PSG. Dataset previously described[23] and obtained via the National Sleep Research Resource.[56] Version used here: "MESA Commercial" available from NSRR, which includes only the 2068 of the original 2237 participants who consented to allow their data to be used for commercial use. Only data from participants with both PSG staging and concurrent actigraphy were used. Actigraphy data were initially linked to PSG data using a record of PSG/actigraphy overlap from the MESA dataset. This results in 1698 potentially usable participants, each with 1 night of data.
Sleep Restriction	Spectrum (Classic)	6.0.9	Laboratory	24-h	In-Bed	2-2 PM	Training, Tuning, Testing	Laboratory study of sleep restriction from 15 participants over 11 days (10 nights). Of 150 total nights, PSG from 7 nights was missing and unavailable for staging. Protocol for this study has been previously reported.[57]
Sound Sleeping	Spectrum Plus	6.0.9	Laboratory	24-h	In-Bed	12-12 PM	Training, Tuning, Testing	Laboratory study of acoustic stimulation during sleep. Data from 8 participants over 4 contiguous days and nights.[52,58]

MESA, Multi-Ethnic Study of Atherosclerosis; PSG, polysomnography into the training, tuning (or "validation"), and testing of the model, some of which have been included in previous reports. The term "validation" refers to the portion of data held out during model training to monitor out of training sample performance and tune hyperparameters, not to suggest that the approach is "valid" per se.[22] The fifth study, MESA, contributed data primarily from a known in-bed interval and was used to supplement the evaluation of model performance outside of the ~24-hour context. MESA data were not used for model training or tuning, but only supplemental model testing. See Table 1 for descriptions of each study.

The protocols for the four studies used for model training, tuning, and testing were individually approved by the Pennsylvania State University Institutional Review Board, and all participants provided written informed consent. For the MESA dataset, "Institutional Review Board approval was obtained at each study site and written informed consent was obtained from all participants."[23] The present work was completed using deidentified data from these five datasets.

Focus technology: wrist actigraphy

Participants wore an ActiWatch Spectrum actigraphy device (Philips-Respironics, Murrysville, PA) on their nondominant wrist. Two Spectrum models were used across the studies, Classic and Plus, which differ in their underlying accelerometer technology. The Classic model uses a piezoelectric accelerometer, and the Plus model uses a microelectromechanical systems accelerometer. However, Philips-Respironics reports that Plus models are designed to output actigraphy counts that are backwards compatible with Classic models, with an average discrepancy of 1.6 minutes of sleep across a night for devices worn concurrently.[24] The Spectrum device model used within each study is listed in Table 1. Actigraphy data were collected and exported from ActiWare software (version for each study in Table 1; Philips-Respironics, Murrysville, PA) with an epoch length of 30 seconds and a wake threshold of "Medium," corresponding to a weighted moving average activity count of 40. ActiWare uses the Oakley algorithm[3] to classify a given epoch as sleep or wake. These Oakley classifications served as a comparison point for TCN classification performance. Activity counts were used as classifier input, and "Off-Wrist Status" was used to identify epochs during which the device was not being worn. To supplement the Spectrum's off-wrist indicator, which may fail to capture some off-wrist epochs, any contiguous sequence of actigraphy epochs with activity count of 0 that were at-least 2 hours in length were additionally marked as off-wrist. Additional fields (eg, interval classification, event markers, light levels) were not used.

Reference technology: PSG staging supplemented with observation of wakefulness

For all datasets, participants were equipped with an EEG montage compliant with AASM standards[25] during either an in-bed interval or continuously throughout the day. Clinically registered polysomnography technicians (RPSGT) retrospectively staged PSG data in 30-second epochs.

For three of the datasets (Sound Sleeping, Deep Sleeping, and Sleep Restriction), participants were monitored during scheduled waking intervals to confirm wakefulness outside of the PSG staged interval – these observed periods were correspondingly labeled as stage "Wake." For these datasets, the RPSGT indicated the "lights-off" portion of the PSG staging, which was used to label the intervals of data considered "in-bed" and attempting to sleep.

Within the EcoSleep dataset, participants were equipped with continuous portable PSG and slept in their home environment over a period of approximately 42 hours on average. As the EcoSleep dataset was collected in a home environment without RPSGT observation of "lights-off," the primary in-bed interval was inferred from the PSG staging by identifying the longest continuous interval of sleep epochs after disregarding periods of awakening less than 1.5 hours.

For the MESA dataset, only data staged with PSG was used, as wakefulness outside of the PSG interval could not be verified; "in-bed" was labeled according to RPSGT "lights-off" indicators.

Design, study setting, and procedures

Temporal alignment of activity counts and PSG staging

Sleep/wake labels and actigraphy counts were temporally aligned prior to analysis, using a procedure adapted from Marino and colleagues.[6] Details of the alignment procedure are included within the Supplementary Material. Statistics on the percent of records per dataset that could not have the lag procedure applied, along with averaged lag and percent of records with an identified lag of 0, are included in Table 3.

Missing data and rejection of recordings

An epoch of data was considered "incomplete" if the staging used to determine sleep/wake state is missing (eg, a gap between PSG recordings), the PSG epoch is not scorable by the RPSGT, or if the actigraphy device is off-wrist or missing data.

Following temporal alignment between actigraphy and staging, records were removed from model development or evaluation based on their percentage of missing data. Records with 25% or more epochs missing staging, or with 25% or more epochs missing actigraphy counts within the staged interval, were excluded from further analysis. The number of records rejected are indicated within Table 2. The percentage of epochs missing staging and percentage of epochs missing actigraphy for the remaining retained datasets are indicated in Table 3.

Each day or night of data was treated as a time series. To preserve the temporal order within each time series, epochs that were missing staging or actigraphy information were retained during training but masked before computing the loss function (ie, they did not contribute to adapting the weights of the network). Because the predictions for a given epoch also depend on the activity count from leading or following epochs, epochs with missing activity counts also had their activity count set to 0.

Table 2
The number of participants and the number of recordings within each data set

Data set	Participants prior to rejection	Days prior to rejection	Days rejected due to missing staging	Days rejected due to missing actigraphy	Days rejected, due to staging or actigraphy	Participants following rejection	Days following rejection
Deep Sleeping	12	36	0	0	0	12	36
EcoSleep	19	37	10	0	10	17	27
MESA	1698	1698	0	8	8	1690	1690
Sleep Restriction	15	143	16	2	17	15	126
Sound Sleeping	8	32	1	1	1	8	31

MESA, Multi-Ethnic Study of Atherosclerosis

Table 3
Data remaining following recording-level rejection. Standard deviation is indicated in parentheses

Data set	Mean age (SD age)	Gender	Percent of epochs without staging	Percent of epochs within staging without actigraphy counts	Percent of epochs prestaging without actigraphy counts	Percent of epochs poststaging without actigraphy counts	Percent of recordings that could not have actigraphy alignment applied	Average actigraphy lag identified (s)	Percent of records with lag of 0 epochs (no adjustment needed)	Valid epochs per recording	Valid epochs per recording within in-bed interval
Deep Sleeping	50.9 (4.0)	4 M/8 F	0.4 (0.2)	1.9 (2.9)	10.2 (26.7)	33.1 (35.6)	0	− 17.5 (191.7)	36.1	2809.4 (82.9)	1024.2 (14.9)
EcoSleep	77.7 (5.2)	5 M/12 F	5.0 (4.7)	2.3 (3.7)	41.5 (43.5)	3.8 (6.5)	0	− 246.7 (292.1)	37	2415.4 (160.0)	897.4 (117.0)
MESA	69.3 (9.1)	773 M/ 917 F	0.0 (0.0)	0.4 (1.9)	46.8 (32.4)	4.2 (13.6)	0	89.0 (171.9)	49.2	1060.2 (178.1)	951.3 (163.0)
Sleep Restriction	22.3 (2.8)	15 M/0 F	0.9 (1.8)	2.8 (3.9)	6.4 (14.4)	13.7 (16.3)	2.4	− 16.8 (44.5)	63.5	2771.1 (119.4)	892.1 (298.0)
Sound Sleeping	40.2 (5.0)	3 M/5 F	0.2 (0.2)	2.9 (4.4)	7.8 (9.2)	14.6 (24.3)	0	− 41.6 (34.4)	22.6	2789.1 (126.1)	1078.2 (16.4)

MESA, Multi-Ethnic Study of Atherosclerosis

Classifying each epoch of data depends not only on the activity count of that epoch, but also epochs in the past and future, with the exact number depending on the hyperparameters of the TCN model (kernel size, and sequence of dilations). Convolution inputs are typically zero-padded to return an output of equal length. To reduce noise that may be introduced by zero padding, any activity counts +-4 hours outside of the staged region that were available were retained, although lacking these values did not lead to a record being removed from analysis. Table 3 indicates the percentage of activity counts in the 4 hours preceding and following the staged interval that were missing.

Machine learning approach

Data partitions. The sleep/wake classifier was trained, tuned, and tested using *k*-fold cross-validation with 5-folds (Fig. 1). Data partitions were constructed via stratified randomization, with data set as strata. The participants were randomly assigned to partitions. This ensures that data from any participant is not included in both the training/tuning and testing partitions within any given fold. Partition assignment was additionally shuffled to prevent the final fold from systematically containing fewer data sets.

Model structure. The sleep/wake classifier was developed within the TensorFlow (v. 2.6.0) deep learning framework with Keras interface. A TCN,[26] an architecture composed of several stacked temporal convolution layers, was used. This choice was influenced by the favorable performance of TCN on sequence labeling tasks, relative to other deep learning architectures such as recurrent neural networks.[26] In addition, temporal convolution is conceptually similar to the weighted moving sum algorithm used in common actigraphy classifiers, such as the Cole-Kripke,[2] Oakley,[3] and Scripps Clinic.[5]

The model was constructed using a symmetric convolution kernel, in which the predictions for a given epoch depend on the value of activity counts at both past and future epochs. The model was alternatively constructed using convolutions that are causal, only operating on data from current and past epochs, included within the Supplementary Material.

The basic structure of the model is shown in Fig. 2. The inputs to the model consist of a time series of activity counts, and a time series of mask values. Mask values are used to indicate missing timepoints, to allow them to be masked from the loss function. Within each fold, activity counts were scaled to the range of 0-1 by diving by the maximum activity count in that training fold.

Hyperparameter tuning and model training. Hyperparameters related to the model structure or training process were optimized independently for each of the 5 cross-validation folds, using the Keras Tuner[27] hyperparameter optimization package (ver. 1.1.0) with Bayesian Optimization tuner. The hyperparameters to be optimized and their possible values are shown in Table 4.

Within each cross-validation fold, 100 models were trained with different combinations of hyperparameters. Each model was trained with a batch size of 32, for up to 300 epochs to minimize binary cross-entropy loss with Adam optimizer.[28] Training for a given

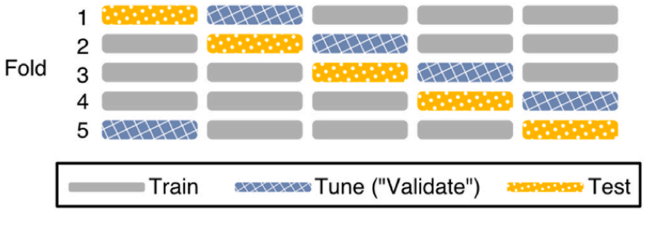


Fig. 1. Structure of train, tune (or “validate”), and test partitions

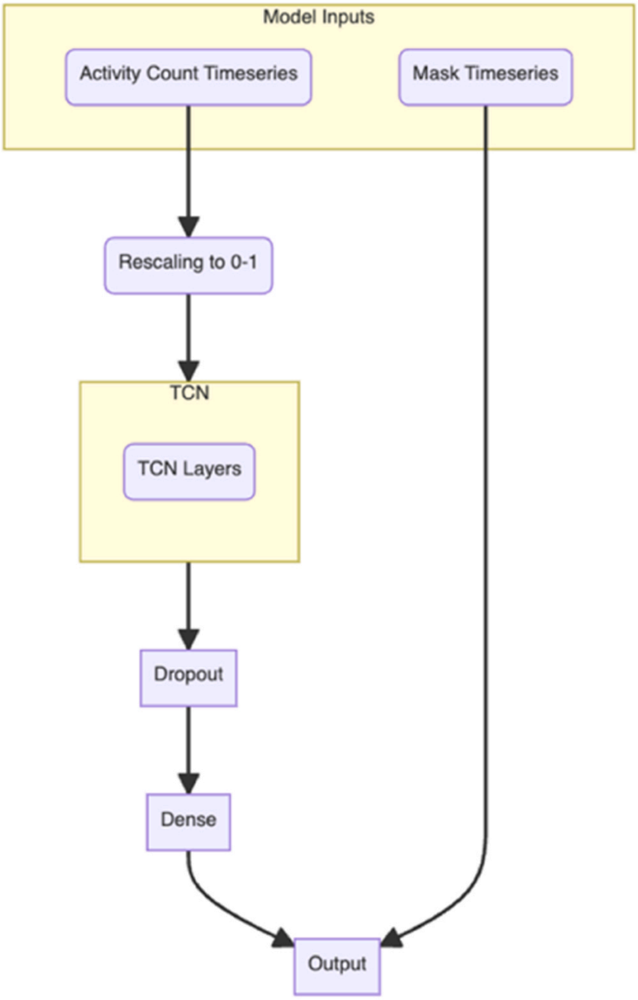


Fig. 2. Model structure. TCN, temporal convolutional network

Table 4
Hyperparameters to tune, with the range of values to search for each

Hyperparameter	Potential values
Number of filters in TCN	1, 2, 4, 8, 16, 32, 64
TCN kernel size	1, 3, 5, 7
Set of TCN dilations	[1, 2], [1, 2, 4], [1, 2, 4, 8], [1, 2, 4, 8, 16], [1, 2, 4, 8, 16, 32]
Dropout rate inside TCN	0.0, 0.1, 0.2, 0.3, 0.4, 0.5
Normalization	layer normalization, batch normalization, none
Dropout rate prior to dense layer	0.0, 0.1, 0.2, 0.3, 0.4, 0.5
Learning rate	0.0001-0.01 with log sampling

TCN, temporal convolutional network.

model was terminated when either 300 epochs had elapsed, or when the loss on the tuning set had not improved for 50 epochs. The weights at the epoch with the lowest tuning set loss were selected for that model. Following a search over 100 combinations of hyperparameter choices, the model with the lowest tuning set loss was selected as the "trained" model for the given fold. The trained model was then applied to the remaining, held-out test data for that fold. The hyperparameter values identified as optimal within each fold of the 5-fold cross-validation procedure are shown in Table 5. The receptive field of the TCN model, the amount of data incorporated into a classification, depends on the kernel size and sequence of dilations. In all folds, the resulting model had a receptive field of 757 epochs, or 3 hours and 9 minutes of data on either side of the epoch being classified.

Classifier evaluation. The primary objective was to determine how well a sleep/wake classifier trained on ~24-hour data could distinguish sleep from wake over a ~24-hour interval. However, to assess how well the ~24-hour model would theoretically perform on known in-bed interval data, results were additionally evaluated on only those epochs that were part of the in-bed interval. Additionally, while the MESA dataset was not used to train the classifier, the trained classifier was tested on the MESA data, both for all-available staged data (which includes, on average, approximately 1 hour of extra data in addition to the in-bed interval), and for the data restricted to the in-bed interval.

For performance reference, the ActiWare software's Oakley sleep/wake algorithm was compared to ground truth for the same data. The Oakley algorithm was selected as reference as we believe it is the algorithm most commonly used to classify sleep/wake from ActiWatch Spectrum actigraphy devices. To ensure comparable evaluation, TCN classifier and Oakley algorithm were compared only on the epochs with both valid TCN and Oakley classification. Classifiers were evaluated in terms of both epoch-by-epoch performance, and in terms of discrepancy in reproducing interval level sleep statistics such as TST, separately for the ~24-hour and in-bed intervals.

At the level of the epoch, the TCN classifier outputs class probabilities, which are used to compute the area under the receiver operating characteristic curve (AUC). Probabilities were mapped to classifications using a threshold of 0.5. Classifications are used to compute accuracy, sensitivity, specificity, balanced accuracy, positive predictive value (PPV), negative predictive value (NPV), F1-score, Matthews correlation coefficient (MCC), and prevalence-adjusted and bias-adjusted kappa[29] (PABAK). As PABAK is a linear transformation of accuracy, all statistical comparisons are identical to accuracy, however, this statistic was included as it has been suggested to be present within sleep technology performance evaluations.[22] The Oakley classifier outputs discrete sleep/wake classifications rather than probabilities, so AUC was not computed. For metrics with classification counts in the denominator that can be 0, the value was set to 0 as division by 0 is undefined. This may occur for example, if calculating NPV in an instance where the classifier didn't classify any

Table 5
Optimal hyperparameters identified within each fold of the model

Hyperparameter	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Number of filters in TCN	64	64	64	64	64
TCN kernel size	7	7	7	7	7
Set of TCN dilations	[1, 2, 4, 8, 16, 32]	[1, 2, 4, 8, 16, 32]	[1, 2, 4, 8, 16, 32]	[1, 2, 4, 8, 16, 32]	[1, 2, 4, 8, 16, 32]
Dropout inside TCN	0	0	0	0.5	0
Normalization	Layer normalization	Layer normalization	Layer normalization	Layer normalization	Layer normalization
Dropout prior to dense layer	0.3	0	0	0	0.3
Learning rate	1.00E-04	3.86E-04	1.05E-04	1.00E-02	6.40E-04

TCN, temporal convolutional network

epochs as wake. Additional description of these metrics is included in the Supplementary Material (1b).

At the level of the interval, epoch-by-epoch classifications were used to obtain TST within the ~24-hour interval, and sleep onset latency (SOL), TST, SE, and wake after sleep onset (WASO) within the in-bed interval. Definitions are included within the Supplementary Material (1c). The discrepancy of each metric to ground truth values were evaluated in terms of bias, mean absolute error (MAE), and concordance correlation coefficient[30] (CCC). Additional description of these measures is included within the Supplementary Material (1d). The EcoSleep dataset was excluded from comparisons of SOL or SE, as the method labeling in-bed for the EcoSleep data cannot identify presleep wakefulness within the in-bed interval.

Scatterplots and Bland-Altman plots are additionally included to visually compare each classifier to ground-truth reference values. Note that the x-axis plots the reference values, rather than the mean of reference and predicted values.[31] Bland-Altman plots were generated using a mixed-effects approach, using the loa_lme function from the SimplyAgree[32] package (ver. 0.1.2) for R. Proportional bias was evaluated for each plot by fitting a linear-mixed effects model predicting the classifier – reference differences with fixed effects of model intercept and reference value, and random effects of intercept nested within participants. If the fixed-effect of reference value was significant at a .05 level, indicating that the difference depends on the value of the reference, proportional bias was plotted. Otherwise, only a fixed mean bias was plotted. In either case, 95% confidence intervals were plotted around bias and limits of agreement, constructed from parametric bootstrap with 10,000 replicates.

The ~24-hour datasets included data collected over multiple days for each participant. To account for clustering of data within participants, epoch-by-epoch, bias, and MAE results were compared between Oakley and TCN classifiers using linear mixed-effects models. Each evaluation metric was evaluated separately using the lme4[33] package (ver. 1.1.31) for R (ver. 4.2.0), with degrees of freedom estimated via Satterthwaite's method using the package lmerTest[34] (ver. 3.1.3). Models included fixed-effects of intercept and classifier type. Each model was initially evaluated with random effects of intercept and classifier type, grouped within participants. If this model produced a singular fit, it was refit with the same fixed effects but only a random effect of intercept, grouped within participants. Descriptive statistics such as mean and standard deviation are reported both as grand means (computed across days regardless of clustering) and as values derived from mixed-effects models. The latter are obtained by fitting a model separately to each condition with a fixed-effect of intercept, and a random-effect of intercept grouped within participants. Such models have both cluster (participant) and residual standard deviation; the population estimates of each are included. The difference displayed is the condition difference estimated from the mixed-effects model used to compute inferential statistics. This value is similar but not necessarily the same as the difference between the mixed-effect condition means, which are estimated independently.

As each participant contributes multiple days of data, mixed-effect sizes (d) were estimated for each outcome metric by dividing the difference between classifiers by the square root of the sum of random variance components.[35]

CCC was calculated across all days regardless of clustering using via the epiR[36] (ver. 2.0.57) package for R, as well as using the variance components approach to compute a longitudinal repeated measures variation of CCC[37] (CCCLON) using the cccrm[38] (ver. 2.1.0) packages for R. In both cases, differences in CCC/CCCLON between classifiers were computed via nonparametric bootstrap of the difference with 10,000 bootstrap replicates. CCC/CCCLON differences were considered statistically significant if the resulting 95% confidence interval of the difference excluded zero.

Results

Core analytics and main outcome variables

Epoch-by-epoch performance

Epoch-by-epoch performance is indicated within Table 6. At the ~24-hour interval, the TCN model produces favorable epoch-by-epoch performance to the Oakley classifications on nearly all the metrics evaluated, excepting sensitivity which does not statistically differ. When restricting the performance evaluation to only the known in-bed interval, the TCN shows favorable epoch-by-epoch performance on accuracy, NPV, F1-score, Matthews correlation coefficient, and PABAK, while the remaining measures do not statistically differ between the two classifiers.

Fig. 3 displays ROC curves for the TCN classifier, separately for ~24-hour and in-bed evaluation. The ROC curve depicts the trade-off between the true positive rate (sensitivity) and the false positive rate (1 – specificity) as the probability threshold for classification is altered. The model performs more favorably across the ~24-hour interval than the in-bed interval, also reflected numerically by the AUC values in Table 6.

Table 7 displays confusion matrices for the performance of the Oakley and TCN classifiers, at both evaluation intervals. The confusion matrices reiterate the increased specificity for the TCN model at ~24-hour evaluation that had been demonstrated statistically in Table 6. In addition, by summing within columns, the base rates of sleep and wake within each interval can be obtained.

Sleep metric discrepancy performance at ~24-hour and in-bed intervals

Discrepancy of classifiers within the ~24-hour and in-bed intervals are indicated in Table 8 for bias and MAE, and Table 9 for CCC/CCCLON. Across the ~24-hour interval, relative to the Oakley classifications, the amount of TST predicted by the TCN model more closely matches ground truth in terms of bias, MAE, and both variants of CCC. Scatterplots and Bland-Altman plots visualizing the relationship between true and predicted TST across the ~24-hour interval are shown in Fig. 4. Within the in-bed interval, the classifiers have fewer differences in terms of reproducing ground truth. The Oakley classifier performs significantly better in terms of bias in SE or WASO, while the TCN classifier performs significantly better in terms of bias and both variations of CCC in SOL. Scatterplots and Bland-Altman plots visualizing the relationship between true and predicted values for these metrics across the in-bed interval are shown in Fig. 5.

Additional analytics and exploratory analyses

Several additional analyses are included within the Supplementary Material: performance of the trained TCN classifier as applied to the MESA dataset, comparison of the TCN classifier to the Sadeh[4] and Scripps Clinic[5] classifiers, and a variation of the TCN classifier trained with causal convolutions (operating only on the current and prior epochs) that is appropriate for "real-time" use.

Discussion

Main results and implications

A TCN model was developed to classify sleep/wake across a ~24-hour interval without knowledge of in-bed timing, using activity counts as input. Performance within the known in-bed interval was additionally evaluated to determine if the same classifier could be used for sleep/wake classification when the in-bed period is known. At the ~24-hour interval, the TCN classifier showed improved epoch-by-epoch performance relative to the Oakley algorithm with respect to multiple criteria. Additionally, TST across the ~24-hour interval

Table 6
Epoch-level classification performance for both the ~24-hour and in-bed intervals for both the Oakley and TCN classifiers

~24-hour interval epoch-by-epoch evaluation

Metric	Oakley (Grand)	TCN (Grand)	Oakley (Mixed)	TCN (Mixed)	Difference	DF	t	*p*	d	Random effect structure
AUC		0.989 (0.017)		0.985 (0.012; 0.013)						
Accuracy	0.771 (0.090)	0.956 (0.043)	0.757 (0.064; 0.066)	**0.947 (0.028; 0.035)**	0.197	47.00	20.67	< .01	3.06	Slope and intercept
Balanced accuracy	0.820 (0.066)	0.956 (0.041)	0.811 (0.043; 0.051)	**0.946 (0.032; 0.031)**	0.140	40.10	19.30	< .01	2.69	Slope and intercept
Sensitivity	0.950 (0.040)	0.958 (0.061)	0.952 (0.032; 0.022)	0.945 (0.065; 0.039)	− 0.006	34.42	− 0.74	.46	0.09	Slope and intercept
Specificity	0.691 (0.135)	0.954 (0.055)	0.670 (0.095; 0.099)	**0.946 (0.030; 0.047)**	0.263	393.18	31.04	< .01	2.54	Intercept
PPV	0.585 (0.125)	0.913 (0.100)	0.573 (0.074; 0.102)	**0.892 (0.068; 0.077)**	0.326	45.71	25.77	< .01	2.90	Slope and intercept
NPV	0.968 (0.028)	0.982 (0.027)	0.969 (0.019; 0.022)	**0.975 (0.029; 0.018)**	0.010	22.59	3.69	< .01	0.33	Slope and intercept
F1-Score	0.715 (0.094)	0.931 (0.068)	0.708 (0.056; 0.077)	**0.914 (0.051; 0.052)**	0.215	42.29	22.43	< .01	2.64	Slope and intercept
MCC	0.593 (0.120)	0.903 (0.091)	0.579 (0.080; 0.092)	**0.880 (0.066; 0.070)**	0.308	41.54	22.85	< .01	2.93	Slope and intercept
PABAK	0.542 (0.180)	0.912 (0.087)	0.515 (0.128; 0.131)	**0.893 (0.055; 0.070)**	0.393	47.00	20.67	< .01	3.06	Slope and intercept

In-bed interval epoch-by-epoch evaluation

Metric	Oakley (Grand)	TCN (Grand)	Oakley (Mixed)	TCN (Mixed)	Difference	DF	t	*p*	d	Random effect structure
AUC		0.901 (0.119)		0.866 (0.105; 0.075)						
Accuracy	0.892 (0.073)	0.906 (0.085)	0.875 (0.062; 0.049)	**0.878 (0.076; 0.054)**	0.014	385.70	2.89	< .01	0.16	Intercept
Balanced accuracy	0.736 (0.118)	0.742 (0.146)	0.704 (0.087; 0.081)	0.700 (0.108; 0.096)	− 0.005	53.04	− 0.44	.66	0.04	Slope and intercept
Sensitivity	0.950 (0.039)	0.960 (0.058)	0.953 (0.032; 0.022)	0.950 (0.058; 0.039)	− 0.002	32.60	− 0.22	.82	0.03	Slope and intercept
Specificity	0.522 (0.242)	0.524 (0.295)	0.455 (0.182; 0.156)	0.450 (0.219; 0.194)	− 0.013	48.44	− 0.57	.57	0.05	Slope and intercept
PPV	0.929 (0.081)	0.934 (0.080)	0.906 (0.071; 0.052)	0.910 (0.070; 0.050)	0.005	386.28	1.11	.27	0.06	Intercept
NPV	0.528 (0.231)	0.676 (0.262)	0.567 (0.174; 0.164)	**0.675 (0.180; 0.206)**	0.131	40.63	5.60	< .01	0.50	Slope and intercept
F1-Score	0.937 (0.048)	0.944 (0.056)	0.926 (0.042; 0.032)	**0.926 (0.053; 0.035)**	0.007	385.02	2.36	.02	0.13	Intercept
MCC	0.442 (0.171)	0.509 (0.234)	0.421 (0.123; 0.129)	**0.451 (0.172; 0.162)**	0.040	53.41	2.12	.04	0.18	Slope and intercept
PABAK	0.784 (0.146)	0.812 (0.169)	0.751 (0.124; 0.098)	**0.757 (0.151; 0.109)**	0.028	385.70	2.89	< .01	0.16	Intercept

AUC, area under the receiver operating characteristic curve; DF, mixed-effects model degrees of freedom; MCC, Matthews correlation coefficient; NPV, negative predictive value; PABAK, prevalence-adjusted and bias-adjusted kappa; PPV, positive predictive value; t, mixed-effects model t-value for the comparison between classifiers; TCN, temporal convolutional network

Bolded values indicate the more favorable outcomes for comparisons that are statistically significant at an alpha level of .05

AUC not possible for Oakley classifier given discrete classification

Mixed-effects models compared the manufacturer's standard algorithm vs. TCN model classification, with days nested within participants

Grand mean values are averaged across recordings regardless of clustering within participants. The standard deviation of these values is represented in parentheses

Mixed-effect mean values are derived by fitting mixed-effect models separately to each condition with a fixed-effect of intercept, and a random-effect of intercept grouped within participants. Standard deviation values are at the level of the cluster, and residual

Difference is the condition difference estimated from the mixed-effects model used to compute inferential statistics. This value is similar but not necessarily the same as the difference between the mixed-effect condition means, which are estimated independently

For EcoSleep Study, in-bed interval was not set for participants, so nighttime sleep interval imputed from PSG scored start/end of nighttime sleep epochs

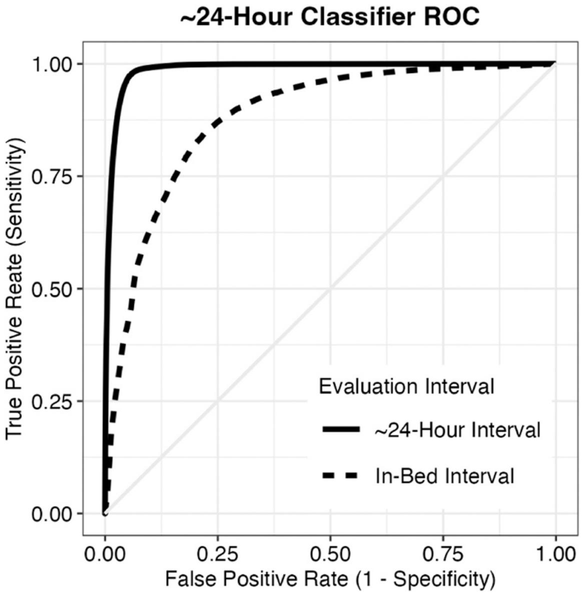


Fig. 3. Receiver operating characteristic (ROC) for classifier performance, collected across the 5 cross-validation folds. Performance is separately displayed for evaluation across the ~24-hour interval (solid line) or only within the in-bed interval (dashed line). Each line is the mean of ROC curves from individual days, combined via threshold averaging.[59] These curves do not account for clustering of days within participants, however mean area under the ROC curve values derived from mixed-effects models accounting for clustering are displayed in Table 6

was better captured by the TCN model. When restricting analysis to the known in-bed interval, epoch-by-epoch performance of the TCN classifier was as good or better than the Oakley classifier. Reproduction of in-bed summary values such as TST, WASO, SOL, SE, was more variable, with the favorable classifier depending on the metric.

Improved classification of sleep/wake from actigraphy counts without knowledge of in-bed timing is relevant for both research and clinical purposes. In both cases, participants may be asked to wear wrist actigraphy devices for several days in their home environment where knowledge of the timing of in-bed intervals may be unknown.

Other works have taken similar approaches, with some key differences. Haghayegh et al[39] also developed a sleep/wake classifier using deep learning techniques, however, this work covered only a known in-bed interval. Palotti et al[40] report sleep/wake classification performance using a range of machine learning approaches, also using the MESA dataset, and report performance both within the in-bed interval and over 24 hours. However, within the MESA dataset, PSG was used in a limited time window primarily restricted to the in-bed interval, with the majority of sleep/wake labeling across the 24-hour interval inferred from researcher annotations (using actigraphy, sleep diaries, and event markers). As these annotations are less precise than PSG at capturing sleep/wake at an epoch-by-epoch level, only data from the MESA dataset with PSG staging was analyzed within our report. Our report also differs by analyzing performance of the ~24-hour classifier at both the ~24-hour and in-bed intervals, to evaluate how performance may change when a model trained across ~24 hours is also used within a known in-bed interval. Jean-Louis et al[41] adapted a classifier developed for in-bed recordings for use in a 24-hour context, evaluating it in both in-bed and 24-hour intervals. This work takes a different approach in that 24-hour use was supported primarily by the addition of rescoring rules following the epoch-by-epoch classifier output.

A separate set of techniques have approached 24-hour classification by first identifying candidate rest intervals that are coarser than the 30-second epoch. Within the Munich Actimetry Sleep Detection Algorithm (MASDA),[42,43] approach, the temporal resolution of epochs under analysis are reduced from the traditional 30 seconds to a coarser 10 minutes. These 10-minute intervals are then compared to a 24-hour moving average and subsequently temporally filtered, to identify consolidated periods of sleep and wake. The MASDA approach is reported to perform particularly well at identifying the timing of sleep intervals, although at the detriment of capturing finer-grained information such as the short periods of awakening within the night that can contribute to WASO.[42] Regalia et al[44] report the evaluation of a two-stage algorithm operating on actigraphy counts termed “ACT-S1,″ which first identifies the time limits of rest intervals, then classifies sleep/wake within the rest intervals. Tudor-Locke and colleagues have also developed a similar approach for waist-worn actigraphy devices that include an inclinometer, reporting good performance at reproducing human-identified sleep intervals in children[45] as well as reproducing sleep logs in adults.[46]

Additional work has approached sleep/wake classification over 24-hour intervals, but using tri-axial accelerometer data (rather than actigraphy counts). The Heuristic algorithm looking at Distribution of Change in Z-Angle (HDCZA)[47] uses arm angle derived from a tri-axial accelerometer to identify the time limits of the rest interval in order to reduce reliance on sleep diary. After the time-window of the rest interval is established, other actigraphy algorithms developed to operate within a rest interval can be used to determine epoch-by-epoch sleep/wake if needed. In another approach, Katori et al[48] report applying a sleep/wake classifier trained on in-bed tri-axial actigraphy data to a large multiday dataset. The sleep/wake output from the classifier was used with rescoring rules to derive a set of sleep quantity and timing indices to cluster participants into sleep phenotypes.

Additional results and implications

Application of the ~24-hour model to the MESA dataset is presented within the Supplementary Material. The ~24-hour TCN model produced more favorable performance on several metrics when analyzing all available data, though several metrics also favored the Oakley classifier. Results were similarly mixed when analysis was restricted to the in-bed interval, though the majority of metrics favored the Oakley classifier. Of note is that the MESA data does not contain staging across 24 hours, but instead includes a limited amount of staging outside of the in-bed interval, meaning the added benefit of a 24-hour algorithm when analyzing all of the available data is more limited. Relatedly, the 24-hour algorithm had less activity count history (valid epochs preceding the staged interval) which may affect the classifier performance for epochs early in the interval. Differences in performance in the in-bed interval between the ~24-hour and MESA datasets may also be attributable to the different average age range, or context in which the datasets were collected. Expanding the data in the training set may generate a model that is more generalizable.

Limitations and future directions

Within each fold of the cross-validation procedure, the set of hyper-parameter values to use were identified from a range of possible values. Values identified were similar across the 5 cross-validation folds, however, several values were at the limit of the searched range; extending the range to search could improve a future model.

The classifier incorporates the activity count in neighboring epochs when making a classification, including 3 hours and 9 minutes on either side of the epoch being classified. To support the classification of epochs on the edges of the staged interval, any valid activity counts that were present outside of the staged interval were

Table 7

Epoch-level confusion matrices for the ~24-hour and in-bed intervals for both the Oakley and TCN classifiers

(a) Confusion matrices for the ~24-h and in-bed intervals computed as "grand" statistics without accounting for clustering of recordings within participants.

	~24-h evaluation (grand mean, SD, CI)		In-bed only evaluation (grand mean, SD, CI)	
	True wake	True sleep	True wake	True sleep
Oakley predicted wake	48.31 (11.78) [46.75 49.87]	01.54 (01.35) [01.36 01.72]	05.05 (04.10) [04.51 05.59]	04.44 (03.45) [03.98 04.89]
Oakley predicted sleep	21.37 (09.26) [20.14 22.59]	28.78 (07.46) [27.80 29.77]	06.38 (07.26) [05.42 07.34]	84.14 (09.65) [82.86 85.41]
	True wake	True sleep	True wake	True sleep
TCN predicted wake	66.54 (08.94) [65.36 67.72]	01.26 (01.92) [01.01 01.51]	05.45 (05.76) [04.69 06.21]	03.43 (04.84) [02.79 04.07]
TCN predicted sleep	03.14 (03.88) [02.62 03.65]	29.06 (07.81) [28.03 30.10]	05.98 (07.28) [05.02 06.94]	85.14 (11.03) [83.68 86.60]

(b) Confusion matrices for the ~24-h and in-bed intervals computed as mixed-effects means to account for clustering of recordings within participants.

	~24-h evaluation (mixed-effect mean, SDs, CI)		In-bed only evaluation (mixed-effect mean, SDs, CI)	
	True wake	True sleep	True wake	True sleep
Oakley predicted wake	46.84 (6.24; 10.05) [44.62, 49.11]	1.48 (0.99; 0.91) [1.18, 1.78]	5.15 (2.01; 3.57) [4.38, 5.91]	4.09 (2.84; 1.82) [3.25, 4.91]
Oakley predicted sleep	22.81 (6.73; 6.53) [20.75, 24.89]	28.78 (0.00; 7.46) [27.80, 29.77]	8.43 (6.42; 4.57) [6.52, 10.34]	82.36 (7.45; 6.90) [80.06, 84.62]
	True wake	True sleep	True wake	True sleep
TCN predicted wake	66.54 (0.00; 8.94) [65.37, 67.74]	1.72 (2.09; 1.22) [1.12, 2.32]	5.57 (3.23; 4.82) [4.42, 6.70]	4.17 (4.55; 3.35) [2.83, 5.53]
TCN predicted sleep	3.73 (2.26; 3.22) [2.95, 4.51]	29.06 (0.00; 7.81) [28.02, 30.07]	8.01 (6.45; 4.46) [6.15, 9.87]	82.25 (9.32; 7.69) [79.46, 85.09]

TCN, temporal convolutional network

Grand mean values represent the percentage of predictions in that cell, averaged across recordings regardless of clustering within participants. The standard deviation and 95% confidence interval of these values are represented in parentheses and brackets, respectively

Mixed-effect mean values are derived by fitting mixed-effect models separately to each condition with a fixed-effect of intercept, and a random-effect of intercept grouped within participants. Standard deviation values are at the level of the cluster, and residual. Note that in some cases the fitted mixed-effect models produced a singular fit with estimated cluster standard deviations near zero. In the mixed-effects case, 95% confidence intervals are computed via percentile bootstrap with 10,000 replicates

Table 8

Discrepancy (bias and MAE) of interval level statistics for ~24-hour and in-bed intervals

~24-h interval evaluation (bias and MAE)

Metric	Measure	Oakley (Grand)	TCN (Grand)	Oakley (Mixed)	TCN (Mixed)	Difference	DF	t	*p*	d	Random effect structure
TST (min)	Bias	272.25 (135.97)	25.69 (59.05)	289.46 (103.96; 90.98)	**30.21 (38.44; 46.97)**	−260.37	48.78	−18.08	< .01	2.81	Slope and intercept
TST (min)	MAE	272.79 (134.90)	39.84 (50.55)	289.75 (103.39; 90.47)	**44.38 (25.40; 44.44)**	−232.95	390.27	−28.30	< .01	2.26	Intercept

In-bed interval evaluation (bias and MAE)

Metric	Measure	Oakley (Grand)	TCN (Grand)	Oakley (Mixed)	TCN (Mixed)	Difference	DF	t	*p*	d	Random effect structure
SE (%)	Bias	1.33 (8.74)	2.48 (8.88)	**3.50 (8.32; 5.02)**	4.25 (8.16; 5.69)	1.15	349.25	2.13	.03	0.12	Intercept
SOL (min)	Bias	−13.17 (22.17)	0.41 (17.03)	−12.37 (10.43; 19.29)	**1.13 (10.35; 14.66)**	13.21	28.40	6.08	< .01	0.68	Slope and intercept
WASO (min)	Bias	1.87 (43.92)	−19.10 (40.03)	**−11.25 (34.44; 29.07)**	−28.99 (34.93; 26.11)	−20.97	386.54	−8.00	< .01	0.47	Intercept
TST (min)	Bias	11.87 (43.71)	14.68 (45.51)	22.54 (37.44; 26.69)	20.13 (40.53; 29.69)	2.80	387.00	1.03	.31	0.06	Intercept
SE (%)	MAE	5.98 (6.49)	5.95 (7.03)	7.07 (5.93; 4.29)	7.34 (6.31; 4.78)	−0.03	348.84	−0.07	.95	0.00	Intercept
SOL (min)	MAE	13.45 (22.00)	9.06 (14.41)	12.73 (10.20; 19.21)	10.00 (8.03; 12.58)	−3.70	28.30	−1.62	.12	0.21	Slope and intercept
WASO (min)	MAE	30.39 (31.69)	28.14 (34.26)	32.39 (20.55; 25.24)	36.20 (29.04; 23.78)	1.26	44.13	0.40	.69	0.03	Slope and intercept
TST (min)	MAE	30.20 (33.70)	31.20 (36.19)	35.48 (26.97; 23.82)	38.26 (29.00; 25.61)	1.00	387.13	0.44	.66	0.03	Intercept

MAE, mean absolute error; SE, sleep efficiency; SOL, sleep onset latency; TCN, temporal convolutional network; TST, total sleep time; WASO, wake after sleep onset

Bolded values indicate the more favorable outcomes for comparisons that are statistically significant at an alpha level of .05

Mixed-effects models compared the Oakley algorithm vs. TCN model classification, with days nested within participants

Grand mean values represent the percentage of predictions in that cell, averaged across recordings regardless of clustering within participants. The standard deviation of these values is represented in parentheses

Mixed-effect mean values are derived by fitting mixed-effect models separately to each condition with a fixed-effect of intercept, and a random-effect of intercept grouped within participants. Standard deviation values are at the level of the cluster, and residual

Difference is the condition difference estimated from the mixed-effects model used to compute inferential statistics. This value is similar but not necessarily the same as the difference between the mixed-effect condition means, which are estimated independently

The EcoSleep dataset was excluded from comparisons of SOL or SE, as the method of imputing in-bed labels for the EcoSleep data cannot identify presleep wakefulness within the in-bed interval

Table 9
Discrepancy (CCC and CCCLON) of interval level statistics for ~24-hour and in-bed intervals

~24-h interval evaluation (CCC and CCCLON)						
Metric	Measure	Oakley	TCN	Difference	Difference excludes 0	Incomplete bootstrap samples
TST	CCC	0.18 [0.14, 0.22]	**0.84 [0.80, 0.87]**	−0.66 [−0.70, −0.61]	*	
TST	CCCLON	0.06 [0.01, 0.11]	**0.64 [0.50, 0.75]**	−0.58 [−0.69, −0.44]	*	20
In-bed interval evaluation (CCC and CCCLON)						
Metric	Measure	Oakley	TCN	Difference	Difference excludes 0	Incomplete bootstrap samples
SE	CCC	0.35 [0.24, 0.45]	0.42 [0.31, 0.53]	−0.07 [−0.16, 0.02]		
SOL	CCC	0.06 [0.02, 0.09]	**0.72 [0.64, 0.78]**	−0.66 [−0.82, −0.45]	*	
WASO	CCC	0.34 [0.23, 0.45]	0.38 [0.29, 0.47]	−0.04 [−0.18, 0.09]		
TST	CCC	0.91 [0.88, 0.93]	0.91 [0.88, 0.93]	0.00 [−0.01, 0.02]		
SE	CCCLON	0.29 [0.07, 0.48]	0.32 [0.12, 0.49]	−0.03 [−0.13, 0.08]		1507
SOL	CCCLON	0.07 [−0.08, 0.22]	**0.65 [0.52, 0.75]**	−0.58 [−0.77, −0.22]	*	896
WASO	CCCLON	0.25 [0.06, 0.42]	0.21 [0.10, 0.33]	0.03 [−0.15, 0.29]		2827
TST	CCCLON	0.77 [0.67, 0.85]	0.76 [0.64, 0.84]	0.02 [−0.02, 0.08]		1

CCC, concordance correlation coefficient, calculated across days without respect to clustering within participants; CCCLON, concordance correlation coefficient with longitudinal repeated measures, calculated across days while accounting for clustering of days within participants; SE, sleep efficiency; SOL, sleep onset latency; TCN, temporal convolutional network; TST, total sleep time; WASO, wake after sleep onset

Asterisks wihin the column "Difference excludes 0" indicate that the 95% confidence interval of the difference excludes 0. In these cases, bolded values indicate the more favorable outcomes

The CCCLON internally fits a mixed-effects model to the data to obtain variance components. Some permutations of the bootstrap can produce data sets that can't be appropriately fit with a mixed-effects model (eg, non-negative approximate variance-covariance). The column Incomplete Bootstrap Samples records the number of instances in which this occurred (out of 10,000)

The EcoSleep dataset was excluded from comparisons of SOL or SE, as the method of imputing in-bed labels for the EcoSleep data cannot identify presleep wakefulness within the in-bed interval

retained, however, a variable amount of extra activity counts were present. While zero-padding is common for deep learning paradigms, the presence of zero-padding at the edges of the staged area may add noise to the classification. Collection of longer data intervals or exploration of alternative padding techniques may improve performance.

The sleep/wake classifier described here uses activity counts as generated by Spectrum actigraphy devices. As there are multiple methods of computing activity counts from the accelerometer time series,[49] and the method used is often proprietary to a given device manufacturer, actigraphy counts have limitations in generalizability, consistency, and transparency. For example, it is not clear how the model presented in this work would perform if applied to actigraphy counts from a non-Spectrum device. These limitations have motivated a shift toward collecting minimally-processed tri-axial accelerometer data rather than activity counts for sleep and clinical research.[50,51] Heart rate, especially as collected by wrist-worn photoplethysmogram devices common on smart watches have also been increasingly used for sleep classification, often in combination with accelerometer data.[52,53] Future work may use these standardized data types.

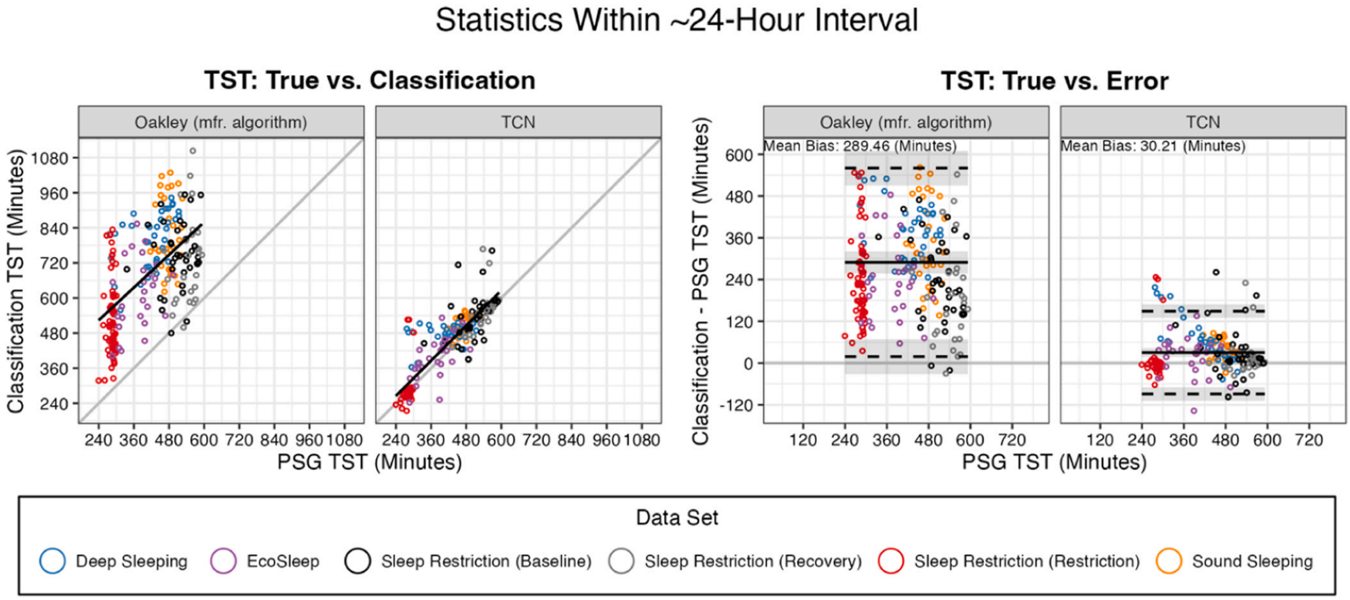
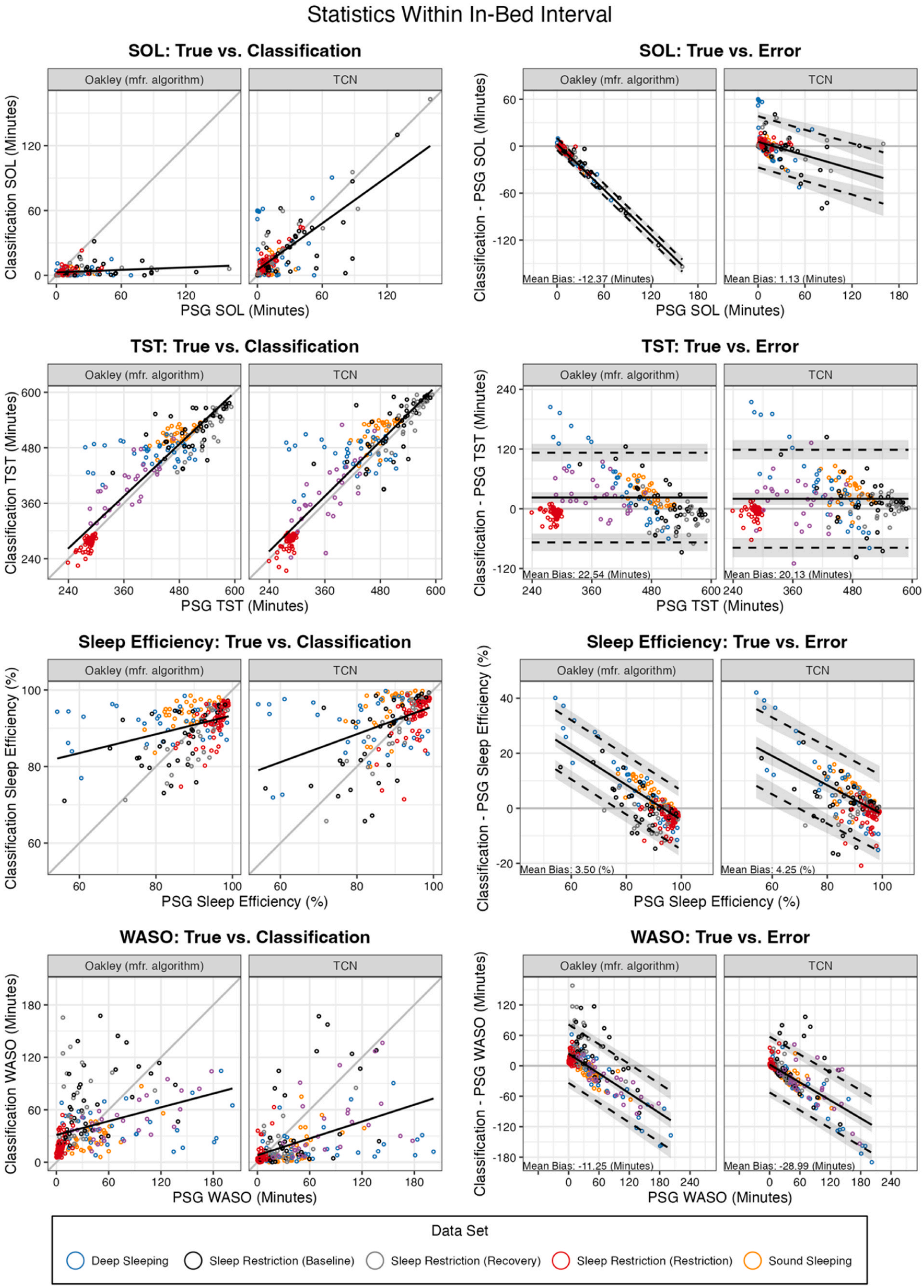


Fig. 4. Polysomnography (PSG)-derived total sleep time (TST) over the ~24-hour interval, compared to the Oakley or temporal convolutional network classifiers. Left column shows PSG vs. classification for each classification type, and for visual reference includes a linear regression of predicted TST on true TST across days (without accounting for clustering within participants; black line). Right column shows Bland-Altman PSG vs. classification error (classification TST - PSG TST), computed using a mixed-effects approach accounting for clustering of days within participants. Proportional bias was evaluated but did not reach statistical significance for either classifier on ~24-hour TST. Indicated are bias (black line) and the upper and lower limits of the 95% limits of agreement (dashed black lines) with 95% confidence intervals for each shaded in gray



(caption on next page)

Fig. 5. Polysomnography (PSG)-derived sleep metrics restricted to the in-bed interval (as defined by registered polysomnography technicians or algorithm in the case of EcoSleep), compared to values from the Spectrum classification outputs or our temporal convolutional network classifier restricted to the same interval. Left column shows the scatterplot of PSG vs. classification derived metric for each classification type, and for visual reference includes the linear regression of predicted metric on true metric (without accounting for clustering within participants; black line). Right column shows a Bland-Altman plot of PSG metric vs. classification metric error (classification metric - PSG metric), computed using a mixed-effects approach accounting for clustering of days within participants. Plotted are sleep onset latency (SOL), total sleep time (TST), sleep efficiency (SE), and wake after sleep onset (WASO). Proportional bias was evaluated and reached significance for all statistics except for TST, which did not reach significance for either classifier. Indicated are bias (black line) and the upper and lower limits of the 95% limits of agreement (dashed black lines), with 95% confidence intervals for each shaded in gray. The EcoSleep dataset was excluded from comparisons of SOL or SE, as the method of imputing in-bed labels for the EcoSleep data cannot identify presleep wakefulness within the in-bed interval

The TCN classifier produced favorable performance to the Oakley classifier when classifying data over ~24 hours. However, the Oakley classifier performed better on some metrics when restricting the analysis to the in-bed interval, for example, bias in SE and WASO. The benefit of using the 24-hour classifier may decrease when the timing of the in-bed interval is known.

Performance may be further improved by providing additional context to the model regarding time of day, circadian phase, or sleep homeostat. As the probability of sleep or sleep stage varies within a rest interval, several sleep/wake or sleep staging classifiers have incorporated time elapsed within the rest interval as a classification feature.[52,54,55] Walch et al[53] expanded this concept by representing temporal context as either a cosine wave rising and falling within a rest interval, or as circadian phase estimated from daytime activity levels. An estimate of an individual's circadian phase or sleep homeostat may particularly benefit 24-hour sleep/wake classification.

Core conclusion

A sleep/wake TCN classifier was trained and evaluated on data labeled over ~24-hour intervals, to eliminate the dependence on knowledge of an in-bed interval. Relative to the commonly used Oakley classifier, the TCN classifier performed statistically equivalent or better on all measures when viewing the data over ~24 hours. When restricting analysis to the in-bed interval, the TCN was still favorable on several metrics. Application of the classifier to the MESA dataset showed decreased performance, especially when restricting analysis to the in-bed interval. Future work should focus on improving generalization performance.

Data and code availability

Preprocessing code, model development code, and trained model files are available at https://github.com/DanielGartenberg/24Hour_Actigraphy_SleepWake.

Funding

SoundSleep and DeepSleep studies were supported via subcontract to Penn State from Proactive Life doing business as SleepSpace on National Science Foundation, USA grant #1622766 (Gartenberg, PI) and National Institutes of Health/National Institute on Aging SBIR, USA R43/44-AG056250 (Gartenberg, PI). Work was conducted at Pennsylvania State University (via subcontract) and further supported by the Penn State Clinical and Translational Sciences Institute, USA (funded by the National Center for Advancing Translational Sciences, National Institutes of Health UL1TR002014) and institutional funds from the College of Health and Human Development of the Pennsylvania State University to Dr. Buxton and Dr. Honavar. The Sleep Restriction study was supported by a pilot grant (PI: Dr. Chang) from the Pennsylvania State University Clinical and Translational Sciences Institute, USA (funded by the National Center for Advancing Translational Sciences, National Institutes of Health, USA through Grant UL1TR002014) and institutional funds from the College of Health and Human Development of the Pennsylvania State University to Drs. Chang and Buxton. Eco-Sleep was supported by funding from the Pennsylvania State University's Social Science Research Institute and a National Institutes of Health, USA grant (U02AG060408) awarded to the Pennsylvania State University's Martin Sliwinski, Center for Healthy Aging. The Multi-Ethnic Study of Atherosclerosis (MESA) Sleep Ancillary study was funded by National Institutes of Health-National Heart, Lung, and Blood Institute Association of Sleep Disorders with Cardiovascular Health Across Ethnic Groups, USA (RO1 HL098433). MESA is supported by National Heart, Lung, and Blood Institute funded contracts HHSN268201500003I, N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute and by cooperative agreements UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420 funded by National Center for Advancing Translational Sciences. The National Sleep Research Resource was supported by the National Heart, Lung, and Blood Institute, USA (R24 HL114473, 75N92019R002). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, National Science Foundation, or the Pennsylvania State University.

Declaration of conflicts of interest

DMR was employed by Proactive Life, Inc. at the time of the initial submission of the manuscript. At the point of revision, he was instead employed by Pennsylvania State University. DG is employed by Proactive Life, Inc., a for-profit company. Related to monitoring sleep, Proactive Life has two patents issued: Sleep Stimulation and Monitoring (US patent 10524661), Cyclical Behavior Modification (US Patent 8468115), four patents pending: Sleep Tracking Method and Device (16/950,987), Valence State Memory Association (16/504,285), Systems, Methods, and Apparatus for Monitoring Sleep (PCT/US21/59978), and two design patents pending: Holder for a Mobile Phone (29/796,124), Stands for Mobile Telephones (WIPO111054). Outside of the current work (last 4 years), Orfeu M. Buxton received honoraria/travel support for lectures/consulting from Boston University, Boston College, Tufts School of Dental Medicine, New York University, University of Miami, University of Utah, University of South Florida, University of Arion, Eric Angle Society for Orthodontists, and Allstate, consulting fees from SleepNumber, and receives an honorarium for his role as the Editor in Chief of Sleep Health (sleephealthjournal.org).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.sleh.2023.07.001.

References

1. Sadeh A. The role and validity of actigraphy in sleep medicine: an update. *Sleep Med Rev.* 2011;15(4):259–267. https://doi.org/10.1016/j.smrv.2010.10.001
2. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. *Sleep.* 1992;15(5):461–469. https://doi.org/10.1093/sleep/15.5.461

3. Oakley NR. *Validation with Polysomnography of the Sleepwatch Sleep/Wake Scoring Algorithm Used by the Actiwatch Activity Monitoring System.* Mini Mitter Co., Inc; 1997.
4. Sadeh A, Sharkey M, Carskadon MA. Activity-based sleep-wake identification: an empirical test of methodological issues. *Sleep.* 1994;17(3):201–207. https://doi.org/10.1093/sleep/17.3.201
5. Kripke DF, Hahn EK, Grizas AP, et al. Wrist actigraphic scoring for sleep laboratory patients: algorithm development. *J Sleep Res.* 2010;19(4):612–619. https://doi.org/10.1111/j.1365-2869.2010.00835.x
6. 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. https://doi.org/10.5665/sleep.3142
7. 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. https://doi.org/10.1093/sleep/26.1.81
8. Paquet J, Kawinska A, Carrier J. Wake detection capacity of actigraphy during sleep. *Sleep.* 2007;30(10):1362–1369.
9. Carney CE, Buysse DJ, Ancoli-Israel S, et al. The consensus sleep diary: standardizing prospective sleep self-monitoring. *Sleep.* 2012;35(2):287–302. https://doi.org/10.5665/sleep.1642
10. Dietch JR, Taylor DJ. Evaluation of the Consensus Sleep Diary in a community sample: comparison with single-channel electroencephalography, actigraphy, and retrospective questionnaire. *J Clin Sleep Med.* 2021;17(7):1389–1399. https://doi.org/10.5664/jcsm.9200
11. Lauderdale DS, Knutson KL, Yan LL, Liu K, Rathouz PJ. Self-reported and measured sleep duration: how similar are they? *Epidemiology.* 2008;19(6):838–845.
12. Jaussent I, Morin CM, Ivers H, Dauvilliers Y. Incidence, worsening and risk factors of daytime sleepiness in a population-based 5-year longitudinal study. *Sci Rep.* 2017;7:1372. https://doi.org/10.1038/s41598-017-01547-0
13. Zhang Z, Xiao X, Ma W, Li J. Napping in older adults: a review of current literature. *Curr Sleep Med Rep.* 2020;6(3):129–135. https://doi.org/10.1007/s40675-020-00183-x
14. Smith MT, McCrae CS, Cheung J, et al. Use of Actigraphy for the Evaluation of Sleep Disorders and Circadian Rhythm Sleep-Wake Disorders: An American Academy of Sleep Medicine Systematic Review, Meta-Analysis, and GRADE Assessment. *J Clin Sleep Med.* 2018;14(7):1209–1230. https://doi.org/10.5664/jcsm.7228
15. Manconi M, Ferri R, Sagrada C, et al. Measuring the error in sleep estimation in normal subjects and in patients with insomnia. *J Sleep Res.* 2010;19(3):478–486. https://doi.org/10.1111/j.1365-2869.2009.00801.x
16. Bianchi MT, Williams KL, Mckinney S, Ellenbogen JM. The subjective–objective mismatch in sleep perception among those with insomnia and sleep apnea. *J Sleep Res.* 2013;22(5):557–568. https://doi.org/10.1111/jsr.12046
17. Frankel BL, Coursey RD, Buchbinder R, Snyder F. Recorded and reported sleep in chronic primary insomnia. *Arch Gen Psychiatry.* 1976;33(5):615–623. https://doi.org/10.1001/archpsyc.1976.01770050067011
18. Edinger JD, Krystal AD. Subtyping primary insomnia: is sleep state misperception a distinct clinical entity. *Sleep Med Rev.* 2003;7(3):203–214. https://doi.org/10.1053/smrv.2002.0253
19. Kay DB, Buysse DJ, Germain A, Hall M, Monk TH. Subjective-objective sleep discrepancy among older adults: associations with insomnia diagnosis and insomnia treatment. *J Sleep Res.* 2015;24(1):32–39. https://doi.org/10.1111/jsr.12220
20. Kanady JC, Drummond SPA, Mednick SC. Actigraphic assessment of a polysomnographic-recorded nap: a validation study. *J Sleep Res.* 2011;20(1pt2):214–222. https://doi.org/10.1111/j.1365-2869.2010.00858.x
21. Depner CM, Cheng PC, Devine JK, et al. Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. *Sleep.* 2020;43(2):zsz254https://doi.org/10.1093/sleep/zsz254
22. de Zambotti M, Menghini L, Grandner MA, et al. Rigorous performance evaluation (previously, "validation") for informed use of new technologies for sleep health measurement. *Sleep Health J Natl Sleep Found.* 2022;8(3):263–269. https://doi.org/10.1016/j.sleh.2022.02.006
23. Chen X, Wang R, Zee P, et al. Racial/ethnic differences in sleep disturbances: the Multi-Ethnic Study of Atherosclerosis (MESA). *Sleep.* 2015;38(6):877–888. https://doi.org/10.5665/sleep.4732
24. Comparison of Sleep Endpoints. Philips Respironics; 2014. accessed 1/27/2023, 3:15:06 PM, Available at: ⟨https://images.philips.com/is/content/PhilipsConsumer/PDFDownloads/Global/Case-studies/HC20191205–001-Actigraphy-Comparison-of-Sleep-Endpts-WhitePaper.pdf⟩.
25. Berry RB, Brooks R, Gamaldo C, et al. *The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications.* American Academy of Sleep Medicine; 2017.
26. Bai S, Kolter JZ, Koltun V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *ArXiv180301271 Cs.* Published online April 19, 2018. Available at: ⟨http://arxiv.org/abs/1803.01271⟩. Accessed February 1, 2021.
27. O'Malley T., Bursztein E., Long J., et al. Keras Tuner. Published online 2019. accessed 7/16/2021, 12:58:49 PM, Available at: ⟨https://github.com/keras-team/keras-tuner⟩.
28. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. Published online December 22, 2014. Available at: ⟨https://arxiv.org/abs/1412.6980v9⟩. Accessed February 7, 2022.
29. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol.* 1993;46(5):423–429. https://doi.org/10.1016/0895-4356(93)90018-V
30. Lin LIK. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989;45(1):255–268. https://doi.org/10.2307/2532051
31. Krouwer JS. Why Bland–Altman plots should use X, not (Y+X)/2 when X is a reference method. *Stat Med.* 2008;27(5):778–780. https://doi.org/10.1002/sim.3086
32. Caldwell AR. SimplyAgree: an R package and jamovi module for simplifying agreement and reliability analyses. *J Open Source Softw.* 2022;7(71):4148. https://doi.org/10.21105/joss.04148
33. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw.* 2015;67(1):1–48. https://doi.org/10.18637/jss.v067.i01
34. Kuznetsova A, Brockhoff PB, Christensen RHB. lmerTest package: tests in linear mixed effects models. *J Stat Softw.* 2017;82(13):1–26. https://doi.org/10.18637/jss.v082.i13
35. Judd CM, Westfall J, Kenny DA. Experiments with more than one random factor: designs, analytic models, and statistical power. *Annu Rev Psychol.* 2017;68(1):601–625. https://doi.org/10.1146/annurev-psych-122414-033702
36. Stevenson M, Nunes ES with contributions from T, Heuer C, et al. epiR: Tools for the Analysis of Epidemiological Data. Published online 2023. accessed 5/9/2023, 3:52:19 PM, R package version 2.0.57. Available at: ⟨https://CRAN.R-project.org/package=epiR⟩.
37. Carrasco JL, Phillips BR, Puig-Martinez J, King TS, Chinchilli VM. Estimation of the concordance correlation coefficient for repeated measures using SAS and R. *Comput Methods Programs Biomed.* 2013;109(3):293–304. https://doi.org/10.1016/j.cmpb.2012.09.002
38. Carrasco J.L., Martinez J.P. cccrm: Concordance Correlation Coefficient for Repeated (and Non-Repeated) Measures. Published online 2022. accessed 5/9/2023, 4:09:18 PM, R package version 2.1.0, Available at: ⟨https://CRAN.R-project.org/package=cccrm⟩.
39. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR. Application of deep learning to improve sleep scoring of wrist actigraphy. *Sleep Med.* 2020;74:235–241. https://doi.org/10.1016/j.sleep.2020.05.008
40. Palotti J, Mall R, Aupetit M, et al. Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *Npj Digit Med.* 2019;2(1):1–9. https://doi.org/10.1038/s41746-019-0126-9
41. 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. https://doi.org/10.1016/S0031-9384(00)00355-3
42. Loock AS, Khan Sullivan A, Reis C, et al. Validation of the Munich Actimetry Sleep Detection Algorithm for estimating sleep–wake patterns from activity recordings. *J Sleep Res.* 2021;30(6):e13371. https://doi.org/10.1111/jsr.13371
43. Roenneberg T, Keller LK, Fischer D, Matera JL, Vetter C, Winnebeck EC. Chapter Twelve - Human activity and rest in situ. In: Sehgal A, ed. *Methods in Enzymology. Vol 552. Circadian Rhythms and Biological Clocks, Part B.* Academic Press; 2015:257–283. https://doi.org/10.1016/bs.mie.2014.11.028
44. Regalia G, Gerboni G, Migliorini M, et al. Sleep assessment by means of a wrist actigraphy-based algorithm: agreement with polysomnography in an ambulatory study on older adults. *Chronobiol Int.* 2021;38(3):400–414. https://doi.org/10.1080/07420528.2020.1835942
45. Tudor-Locke C, Barreira TV, Schuna JM, Mire EF, Katzmarzyk PT. Fully automated waist-worn accelerometer algorithm for detecting children's sleep-period time separate from 24-h physical activity or sedentary behaviors. *Appl Physiol Nutr Metab Physiol Appl Nutr Metab.* 2014;39(1):53–57. https://doi.org/10.1139/apnm-2013-0173
46. Barreira TV, Redmond JG, Brutsaert TD, et al. Can an automated sleep detection algorithm for waist-worn accelerometry replace sleep logs. *Appl Physiol Nutr Metab.* 2018;43(10):1027–1032. https://doi.org/10.1139/apnm-2017-0860
47. van Hees VT, Sabia S, Jones SE, et al. Estimating sleep parameters using an accelerometer without sleep diary. *Sci Rep.* 2018;8(1):12975. https://doi.org/10.1038/s41598-018-31266-z
48. Katori M, Shi S, Ode KL, Tomita Y, Ueda HR. The 103,200-arm acceleration dataset in the UK Biobank revealed a landscape of human sleep phenotypes. *Proc Natl Acad Sci USA.* 2022;119(12):e2116729119. https://doi.org/10.1073/pnas.2116729119
49. 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. https://doi.org/10.1093/sleep/26.3.342
50. Bai J, Di C, Xiao L, et al. An activity index for raw accelerometry data and its comparison with other activity metrics. *PLoS One.* 2016;11(8):e0160644. https://doi.org/10.1371/journal.pone.0160644
51. Bai J, He B, Shou H, Zipunnikov V, Glass TA, Crainiceanu CM. Normalization and extraction of interpretable metrics from raw accelerometry data. *Biostatistics.* 2014;15(1):102–116. https://doi.org/10.1093/biostatistics/kxt029
52. Roberts DM, Schade MM, Mathew GM, Gartenberg D, Buxton OM. Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography. *Sleep.* 2020;43:zsaa045. https://doi.org/10.1093/sleep/zsaa045
53. 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.* 2019;42(12):zsz180https://doi.org/10.1093/sleep/zsz180
54. Redmond SJ, de Chazal P, O'Brien C, Ryan S, McNicholas WT, Heneghan C. Sleep staging using cardiorespiratory signals. *Somnologie Schlafforschung Schlafmed.* 2007;11(4):245–256. https://doi.org/10.1007/s11818-007-0314-8
55. Fonseca P, den Teuling N, Long X, Aarts RM. A comparison of probabilistic classifiers for sleep stage classification. *Physiol Meas.* 2018;39(5):055001. https://doi.org/10.1088/1361-6579/aabbc2

56. Zhang GQ, Cui L, Mueller R, et al. The National Sleep Research Resource: towards a sleep data commons. *J Am Med Inf Assoc.* 2018;25(10):1351–1358. https://doi.org/10.1093/jamia/ocy064

57. Ness KM, Strayer SM, Nahmod NG, Chang AM, Buxton OM, Shearer GC. Two nights of recovery sleep restores the dynamic lipemic response, but not the reduction of insulin sensitivity, induced by five nights of sleep restriction. *Am J.* 2019;316(6):R697–R703. https://doi.org/10.1152/ajpregu.00336.2018

58. Schade MM, Mathew GM, Roberts DM, Gartenberg D, Buxton OM. Enhancing slow oscillations and increasing N3 sleep proportion with supervised, non-phase-locked pink noise and other non-standard auditory stimulation during NREM sleep. *Nat Sci Sleep.* 2020;12:411–429. https://doi.org/10.2147/NSS.S243204

59. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett.* 2006;27(8):861–874. https://doi.org/10.1016/j.patrec.2005.10.010