A Personalized and Adaptive Distribution Classification of Actigraphy Segments into Sleep-Wake States

Austin Vandegriffe[1*], V.A. Samaranayake[1], Matthew Thimgan[1*]

1 Missouri University of Science & Technology, Rolla, Missouri, United States of America

* co-corresponding authors: austin.g.vandegriffe@gmail.com; thimgan@mst.edu

Abstract

Wearable actimeters have the potential to greatly improve our understanding sleep in natural environments and in long-term experiments. Current technologies have served the sleep community well, but they have known weaknesses that introduce errors that can compromise reliable and relevant clinical and research sleep and wakefulness profiles from these data. Newer data collection technologies, such as microelectromechanical systems (MEMS), offer opportunities to gather movement data in different forms and at higher frequencies, making new analytical methods possible and potentially advantageous.

We have developed a novel statistical algorithm, called the Wasserstein Algorithm for Classifying Sleep and Wakefulness (WACSAW), that is based on optimal transport statistics and uses MEMS data as its input. WACSAW segments group data into periods with similar movement patterns and uses optimal transport methods to generate a Wasserstein profile for each segment. The second utilization of optimal transport methodology measures the difference between each segment profile and a hypothetical segment of idealized sleep. Characteristic functions, derived from individual activity segments, were clustered and classified the segment as sleep or wakefulness. WACSAW was initially developed on a 6-person cohort and applied to an additional 16 independent participants. WACSAW returned >95% overall accuracy in sleep and wakefulness assignments validated against participant logs. Compared to the Actiwatch Spectrum Plus, WACSAW delivered a ∼10% improvement in accuracy, sensitivity, and selectivity and showed a reduced standard error between participants, indicating WACSAW conformed to individualized data. In addition, we directly compared WACSAW to GGIR, a current used algorithm designed to accept MEMS data. WACSAW showed an improvement in overall statistics and handled the time series and segmentation differently, which may contribute unique information to activity recordings. Here, we provide novel statistical approach to actimetry that improves sleep/wakefulness designations, adapts to individuals, provides interim metrics that further interpretations and is open source for modification.

Author summary

Wearables are an emerging class of technologies that have the potential to provide real-time, important information to individuals based on their unique biological and behavioral makeup. For 40 years, actimetry has been applied to identifying sleep and wakefulness in natural living environments as sleep changes between laboratory and home environments. Yet, these valuable analyses have weaknesses that may compromise

accuracy. Newer actimetry wearables can collect high frequency movement data and offer an opportunity for different types of analyses. We have developed a novel algorithm called the Wasserstein Algorithm for Classifying Sleep and Wakefulness (WACSAW). WACSAW employs optimal transport statistics to compare movement variation between different time segments to classify sleep or wakefulness for that segment. WACSAW produces >95% accuracy across the 24-hr day for both sleep and wakefulness categorization and more accurately classifies behavior than the Actiwatch Spectrum Plus. In addition, WACSAW output has interim metrics that can be used to assess reliability of the output and requires no human intervention to run. WACSAW may help achieve observations in daily living situations to determine factors that alter sleep, understand the variation in sleep, and set the stage for more home diagnoses and disease identification.

Introduction

Inadequate sleep is being increasingly linked to numerous harmful diseases. Insufficient sleep has been linked to 7 of the 15 leading causes of death in the U.S. [1] and is highly correlated to diabetes [2, 3], cardiovascular disease [4–6], cognitive [7], inflammation and immune dysfunction [8–10], and among other possible health complications, including earlier death [5, 11–15]. Many of these studies are based on self-report and subjective sleep reporting can be prone to errors [16, 17]. Unfortunately, sleep deprivation continues to be a pervasive problem with sleep time potentially decreasing and individuals choosing other activities over sleep [18, 19] which can reduce the overall public health. One confound to these laboratory approaches is that people sleep differently in real world environments than in a lab setting [20] and populations in which these effects have been demonstrated are typically kept small to accommodate laboratory space and time. Thus, objective recording in real world environments would provide practical insight into how natural sleep varies across individuals, manifests in clinical populations, and relates to increased likelihood of health disorders.

Wearables, such as wrist actimeters, have been employed by the sleep field for the last 40 years to attempt to gather objective sleep data in the natural environment. They have emerged as a practical method for determining sleep parameters, especially in long-term or field experiments where polysomnography (PSG), the gold standard for sleep detection, is impractical. In 2007, the American Academy of Sleep Medicine determined that actigraphy was accurate enough to determine total sleep time for both healthy individuals and individuals with sleeping disorders [21]. Thus, there is a general interest in having an automated system that can be deployed widely across the population and does not require a large expense of human effort.

Despite the successes of current actimetry, there remain acknowledged weaknesses in current data collection and analysis that limit the utility of these algorithms. Two of the most commonly employed approaches are the Sadeh [22] and Cole-Kripke [23] algorithms were developed using regression methods. While these approaches have provided valuable insights into sleep biology, regression methodologies optimizes results for the population statistics while some individuals may substantially deviate from the overall average, which distorts the reliable interpretation of every individual without an objective way to validate sleep and wakefulness assignments. Additionally, these regression-based algorithms have difficulty identifying quiet wakefulness [24], possibly due to sampling rate, computational or statical power. On the whole, these algorithms have provided useful data for understanding variations in naturalistic sleep and these approaches continue to provide data today. Yet, technological innovations in sensor, computer, and statistical technology allow for improvements in sleep and wakefulness categorization from movement data.

More recent algorithms have applied machine learning techniques [25–27], where a major focus has been on potentially determining sleep stages from actimetry. It seems that more sophisticated algorithm were not any more accurate than average regression-based algorithm results. Van de Water and collegues [24] applied an AI-based algorithm that yielded behavioral categorization with 86-99% accuracy for sleep sensitivity (i.e., the ability to detect sleep) but only a 27-68% accuracy for sleep specificity (i.e. ability to detect wake), likely due to misclassification of quiet wakefulness activities, which includes desk work, watching television, and reading [28]. When the performance of numerous popular algorithms were directly compared, the Sadeh algorithm was best for healthy individuals, even compared with AI-based methods [29]. The machine learning algorithms may suffer from a high sensitivity to the training set that does not generalize to a novel set of individual data [30–32]. Thus, there remains a need for an improved actimetry algorithm that will be more automated, facilitate 24-hr recording and begin to adapt to individual activity profiles. In addition, other attempts, such as GGIR , have integrated new data collection methods to address questions of interest in sleep [33, 34].

Here we present a novel algorithm, the Wasserstein Algorithm for Classifying Sleep and Wakefulness (WACSAW), that employs a statistical and data-driven approach to classify those segments into sleep and wakefulness states based on raw activity data. WACSAW applies optimal transport and clustering methods to first segment movement profiles and then to determine sleep or wakefulness. WACSAW makes use of high frequency movement data collected using MEMS actimetry devices. When 24 hr data were compared to a detailed activity log, WACSAW correctly categorized >95% of the states. We also compared WACSAW's performance with the commercial Philips Respironics' Actiware 6.0.9 algorithm on Actiwatch Spectrum Plus data [35], as well as a second comparison to GGIR [33]. WACSAW is shown to perform better during quiescent activities and generates personalized results and minimizes both intra- and inter-individual variability of the results. With a median accuracy, sensitivity, and specificity >95% subject to detailed activity logs, WACSAW demonstrates the feasibility of a statistical approach to actigraphy. Thus, WACSAW is a novel algorithm that takes in higher frequency data and returns a more accurate categorization of sleep and wakefulness. Additionally, it is adaptive to individuals, contains explainable parameters, has interim metrics that can be used to ensure reliability of performance within an individual and may help in identifying other conditions within sleep based on movement recordings.

Materials and methods

Participants

Participants were recruited by word of mouth and targeted emails. Study coordinators presented Informed Consent information, enrolled participants, and provided instructions to conduct the study. Upon enrollment, participants completed a demographic survey, including deliberate confirmation that they have no diagnosed sleep disorders and were provided a GENEActiv Original (Activinsights Ltd., Kimbolton, UK) wrist-worn activity monitor [36]. The protocol was approved by the University of Missouri Institutional Review Board (IRB 2026643).

Participants across all cohorts were aged 18-72 years (average 37.7±14.4 y.o), 55% males, and 77% White (Table 1). We implemented two independent protocols: (1) wear the GENEActiv alone and keep a detailed log of activities and other conditions, such as light levels and sleepiness or (2) simultaneously wear the GENEActiv and Philips Respironics Actiwatch 2 (Cambridge Neurotechnology Ltd., Cambridge, UK and Mini

Mitter, Respironics Inc., Bend, Oregon, USA) for two days and keep a similar detailed activity log as in the first protocol. There was a subset of individuals that participated in both protocols as well as some that only completed one or the other. Participants were instructed to wear the watch(es) on their nondominant wrist 24 hours a day. The activity log was completed in either a Qualtrics survey accessed through a mobile device or by maintaining an Excel spreadsheet. Participants were asked to reflect on activities at least 2 times per day but encouraged to report as often as possible to increase accuracy in the reports. Surveyed data included start time of activity, end time of activity, activity level, alertness, and ambient light conditions. Temperature and activity measurements from the GENEActiv actimeter were used to validate watch-on compliance.

Participant #	Trial #	Age	Sex	Ethnicity
WACSAW development cohort (n=6)				
1	1	24	F	White
2	1	47	M	White
3	1	23	M	White
4	1	72	M	South Asian
5	1	26	M	South Asian
6	1	60	M	White
WACSAW independent validation cohort (n=16)				
7	1	47	F	White
8	1	21	M	White
9	1	25	M	White
10	1	30	M	East Asian
11	1	40	F	White
12	1	25	F	East Asian
13	1	21	F	White
14	1	49	M	Hispanic
15	1	23	M	White
16	1	27	F	White
17	1	26	M	White
18	1	50	F	White
19	1	18	M	White
20	1	18	M	White
21	1	45	F	White
22	1	47	M	White
Actiwatch comparison cohort (n=15)				
1	2	24	F	White
2	2	47	M	White
3	2	23	M	White
4	2	72	M	South Asian
5	2	26	M	South Asian
7	2	47	F	White
9	2	25	M	White
10	2	30	M	East Asian
16	1	27	F	White
17	1	26	M	White
18	1	50	F	White
19	1	18	M	White
20	1	18	M	White
21	1	45	F	White
22	1	47	M	White

Table 1. Participant demographics table.

Validation of WACSAW classification

Movement data from the GENEActiv watch was collected at 10 Hz frequency and downloaded with the ActiveInsights software. Force measures were gravity compensated according to manufacturer specifications. After WACSAW classification, performance was compared against participant activity logs using 1 second epochs. Validation metrics included: Accuracy (% of correct classifications over a given time period); Sleep Sensitivity (% correctly classified sleep over the total labeled sleep); and Sleep Specificity (% correctly classified wake over the total labeled wakefulness).

One concern with detailed activity logs is that they can be subjective and reflect an estimate of start and stop rather than exact times of activity transitions [37]. The human error can impact the validation metrics. Thus, we compared WACSAW output to participant logs using two methods. The first uses the raw activity log data provided by the participant (referred to as 'Raw' data comparison). The second involved a research assistant to manually adjust the start and end times subject to the following criteria: there was a sudden change in volatility of a participant's movement within ± 20 minutes of the participant indicating there was a change in activity. In such circumstances, the start and endpoint of the activity were adjusted to this timepoint. Adjustments were also made if a participant noted an activity during a reported interval, but no precise time was given. For example, if they stated they got up to check on an animal in the night without providing a time, we assumed a small period of movement during that night represented their logged activity. We refer to these data as adjusted log or adjusted data (Adj), and these data are analyzed independently from the raw data. Both sets of data are presented in this report, though there are only minor differences.

Comparison to GGIR

Our study includes a comparison with the method introduced by van Hees et al. [38], designed to classify sedentary behavior. This method operates on a time series of tilt angles, which represent movement activity. Posture changes are defined as instances when the tilt angle exceeds a certain threshold, indicating a significant shift in the wearer's position. The van Hees method uses predefined time and angle thresholds to identify these posture changes. This process creates temporal clusters of activity. Whenever two significant posture changes occur within a time period shorter than the predefined time threshold, they are considered part of the same active period, and if these shifts are closer in time than the specified threshold, the segments are merged together, akin to linking activity segments and indicating continued activity. A posture shift is defined based on the instantaneous intensity of the movement, contrasting with WACSAW, which analyzes local volatility. When more than one such posture change is detected, the method identifies periods where these changes occur less frequently than the specified time threshold. If such periods exist, they are classified as periods of no significant posture change, implying sedentary behavior. The selection of time and angle thresholds are crucial hyperparameters in this approach. van Heese et al. [38] found that a time threshold of 5 minutes and a tilt angle threshold of 5 degrees generally yielded the most reliable results. For our comparison, we have used these same values [38].

Development and details of WACSAW

Our goal was to develop an algorithm that improves sleep and wakefulness assignments from raw movement data. We felt we could improve accuracy because new MEMS technology can gather data more frequently and we had access to the raw data with the

GENEActiv actimeter [36]. Moreover, we reasoned that our approach may provide an explainable sleep-wake categorization that may help researchers and those with sleep disorders, may be able to be tuned to an individual, and may provide interim metrics that could increase the information that could be used as a reliability metric. We were not aiming to identify the stages of sleep, such as those detected by polysomnography (PSG). For these reasons, we felt a new algorithm may add new information to interpreting movement data.

We initially developed our algorithm using the movement data from 6 individuals (Cohort #1, Table 1). Because of the limited sample size of participants, we wanted to introduce participants that appeared to have sleep and wakefulness patterns that posed known hardships to past algorithms. This cohort spanned a large age range from 18-72 (ave 42 ± 19.1 y.o.; 83% male; 66% white). Importantly, we developed the algorithm with participant data that contained low activity wakefulness data, for example reading in bed or watching TV, at least one instance of split sleep at night, and significant nighttime movement which has posed problems for past algorithms. We reasoned that if the algorithm could correctly classify these events, then it would likely generalize better. From the detailed participant logs, we chose to develop the algorithm on participants that reported (1) split nighttime sleep; (2) very low overall movement throughout the day and during waking activities; (3) older individual with more movement during sleep; (4) young with sporadic sleep patterns; (5) female with nighttime awakenings; and (6) significant nighttime movement that may be similar to numerous sleep disorders. All criteria were determined by eye and were included to ensure that we were training on activity records with challenging situations. These participants were used for all development sections and the algorithm was validated on a larger, independent cohort after development.

Step 1: Data preparation

The GENEActiv device uses MEMS to record component of gravitational force changes in perpendicular x, y, and z directions where x and y are in the plane of the face of the watch and z is perpendicular to that plane. Data were recorded at 10 Hz frequency resulting in a representative movement time series (Fig. 1). It has been determined that movement away from the direction of gravity sufficiently captures meaningful movement in all 3 directions and can be captured with the tilt angle time series [28]. The tilt angle reduces the dimension of data stream needed for developing the algorithm. Thus, data from the 3 directions were merged to determine the tilt angle (θ_t), by the following formula (Eq. 1):

$$\theta_t = \arccos\left(\frac{z_t}{\sqrt{x_t^2 + y_t^2 + z_t^2}}\right), \tag{1}$$

where x_t, y_t, and z_t are the force changes in the x, y, and z axes at time (t) (Fig. 1B). The tilt angle preserves areas of high and low volatility. In addition, there are periods that are in between high and low volatility (orange label, Fig. 1) that represents quiet wakefulness according to the activity logs.

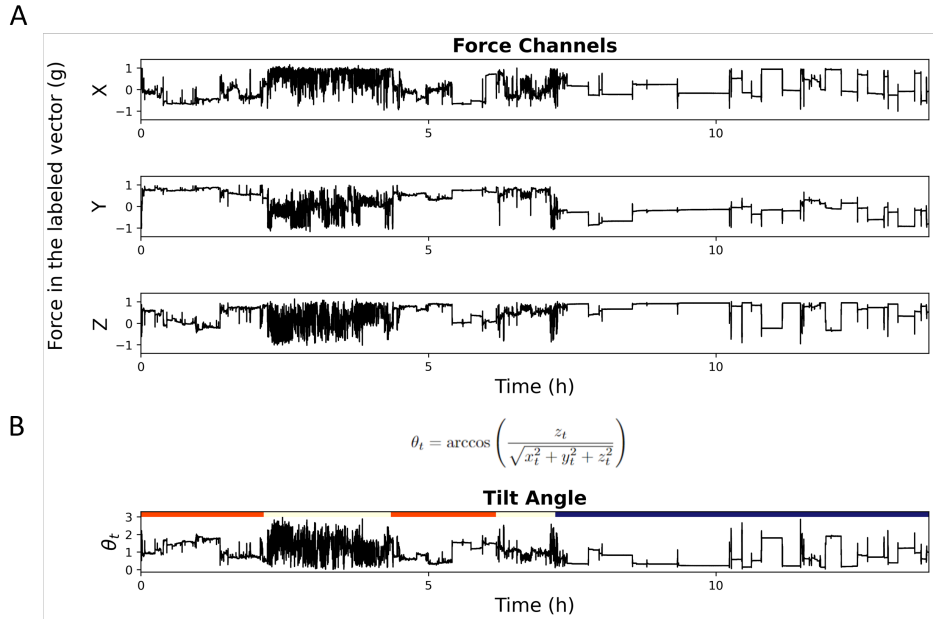


Fig 1. Conversion of raw data to a tilt angle time series. (A) Fifteen hours of raw time series force recordings for one individual from the GENEActiv wrist actimeter. The top 3 time series are the raw recordings of movement in the labeled X, Y, and Z directional vectors. (B) Force changes in the z direction from A is converted into the tilt angle (θ_t) using the presented formula. The spans at the top of the bottom figure indicate wakefulness (white), periods of low activity wakefulness (orange), and sleep periods (blue), as assigned by the participant's activity log.

Step 2: Segmentation

Higher frequency data captured by the MEMS sensors provide more opportunity to detect rapid changes in the tilt angle (Fig. 2A). To isolate changes in the tilt angle irrespective of the absolute magnitude of force, we took the first difference of the tilt angle ($\Delta\theta_t$; Eq. 2), which is the difference between adjacent observed angles θ_t and θ_{t-1} (Fig. 2B):

$$\Delta\theta_t = \theta_t - \theta_{t-1}. \tag{2}$$

Periods of high activity (and thus high volatility) will produce distributions of first differences with larger magnitude compared to those periods of lower activity. Notably distributions of the first difference with varying characteristics can be observed. For instance, participant logs confirm different activity periods, such as high (walking, Fig. 2C), intermediate (reading in bed, Fig. 2D), and low (sleeping, Fig. 2E) in which the high active periods have distributions with a higher standard deviation compared to the quiescent and sleep period distributions. However, the difference in reading in bed (Fig. 2D) and sleeping (Fig. 2E) is subtle, with quiescent wakefulness exhibiting a noticeably higher variability relative to a validated sleep segment. To capture and emphasize such differences and to distinguish between states, we need a sensitive, yet robust method for comparing distributions. Going forward, we will refer to the first difference of the tilt angle as "movement" and the distribution of the first difference of the tilt angle as "movement distribution."

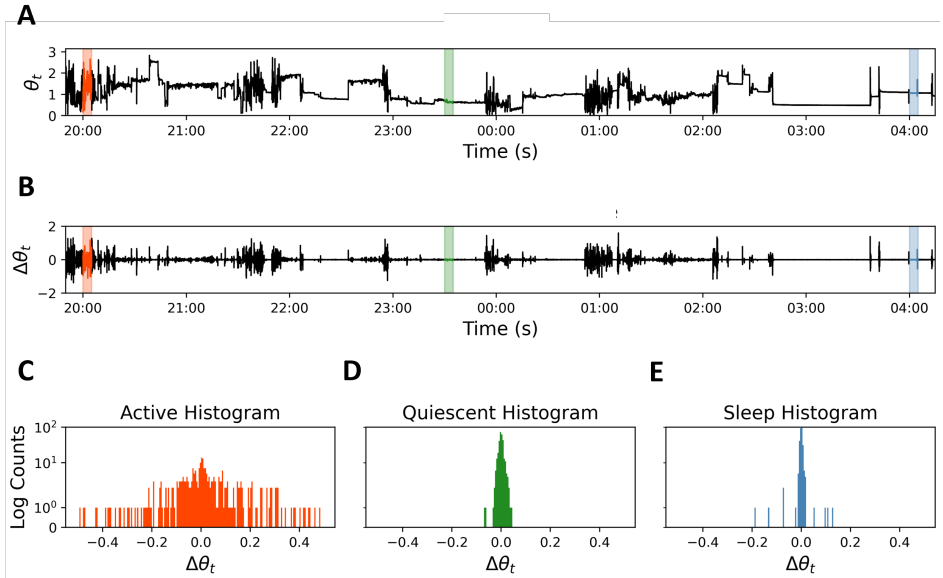


Fig 2. Distribution of movement variability during different activity intensities. (A) The tilt angle time series from 8 hours of activity recordings. These recordings have both a variation and intensity component. (B) First difference (FD) conversion of adjacent points of the raw tilt angle series from A ($\Delta\theta_t$) isolates the movement variability from the intensity level. The orange background indicates an intense period of movement, the green indicates more quiescent wakefulness, and blue indicates sleep according to activity logs. (C-E) The first differences within these 5 min periods were plotted as a histogram in the corresponding colors to visualize the differences in movement. Note the similar nature of quiescent wakefulness (D) and sleep (E).

The Wasserstein Time Series and Segmentation

The use of the Wasserstein metric, a scalar measure based on optimal transport theory, is well suited to distinguishing subtle distributional changes. The greater the difference between two distributions are from one another, the higher the Wasserstein distance between them. Mathematically, the Wasserstein distance is given by (Eq. 3)

$$\mathcal{W}^{(p)}(\boldsymbol{x}, \boldsymbol{z}) = \left(\frac{1}{n} \sum_{t=1}^{n} |x_k - z_k| \right)^{\frac{1}{p}}. \tag{3}$$

where $\boldsymbol{x}$ and $\boldsymbol{z}$ are two time periods of the same length n, with the first differences for that time period sorted in ascending order. We use the notation $\boldsymbol{x}_k$ and $\boldsymbol{z}_k$ to denote the kth element of the sorted periods ($x_k \leq x_{k+1}$; $z_k \leq z_{k+1}$), and p to denote the selected order of the Wasserstein distance. Note that in the current context, the elements $\boldsymbol{x}_k$ and $\boldsymbol{z}_k$ denote the ordered k^{th} first difference tilt angle ($\Delta\theta$) corresponding to the two time periods that are compared. Details about the order p will be provided in a later section (see "Step 3: Classification"). A detailed treatment of optimal transport (OT) is presented in [39, 40] and methods of computational OT in [41].

The Wasserstein distance, in our use case, converts differences between two movement distributions into a single score (Fig. 3). Given a 1 second point of interest (Fig. 3A), WACSAW compares the movement distribution of the 5-minutes interval prior to the point of interest (arrow) to the reference time point against that of the 5-minute interval following the reference point (Fig. 3B). In Fig. 3A, we present a 20 minute

example of movement and its distributions in an abrupt transition from wakefulness to sleep. The aforementioned procedure is applied second to second across the time series generating a Wasserstein score for each second comparing the previous and future distributions, this resulting series is called the Wasserstein Time Series (WTS).

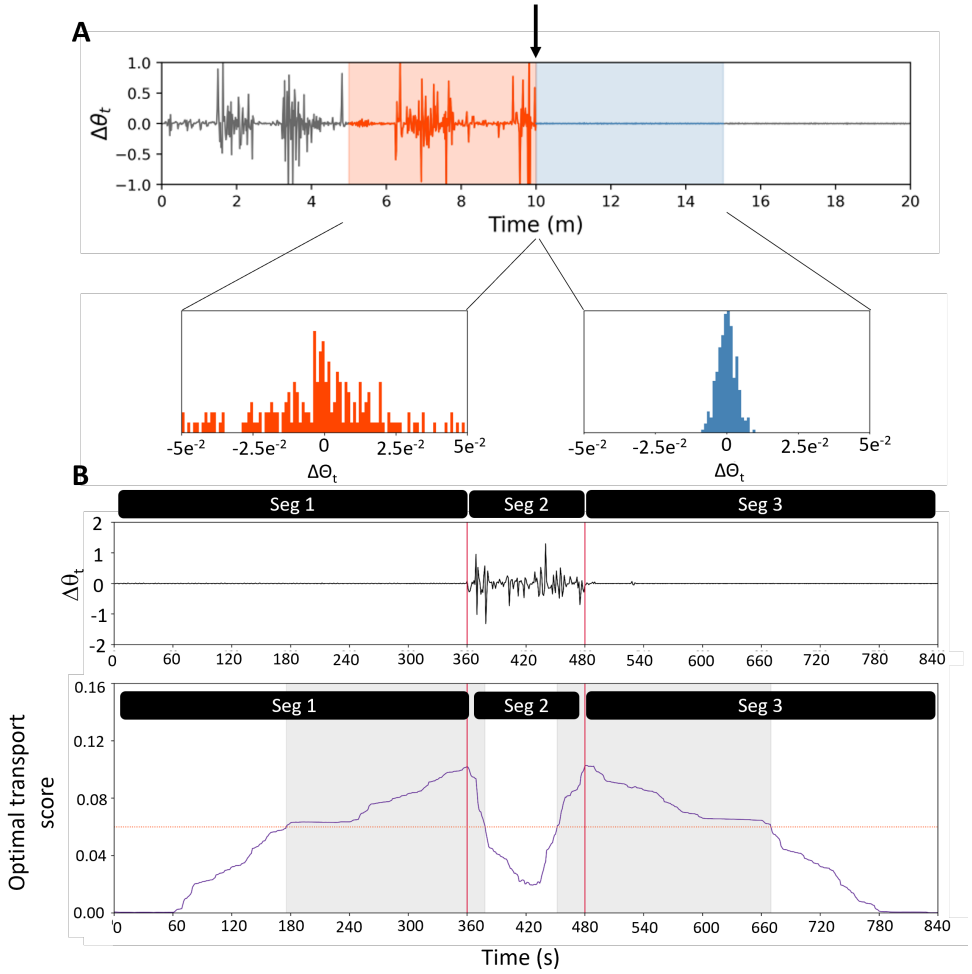


Fig 3. Example of the strategy for obtaining segments within an activity time series. (A) To determine the optimal transport score for a given point (vertical arrow), we compared the distribution of the FD of the tilt angle for each second from the 5 minutes prior (orange shaded area) to the distribution of FD in the 5 minutes after the point of interest (blue shaded region). The distributions for this example are shown in the bottom half of the panel. (B) In a magnified example of where the segmentation points are selected, there is a quiet period (Seg 1) next to a short active period (Seg 2), and then the participant returns to a quiet period (Seg 3). Because the quiet period of Seg 1 is longer than 5 min, there is a very small score at the beginning of the Wasserstein scores. As the WACSAW calculation moves into a region of larger difference, the Wasserstein score increases and crosses a threshold (horizontal orange stippled line; see text for details on how it is calculated). The whole area above the threshold is designated with a gray background. From this region, the peak of the Wasserstein score is identified and marked as the changepoint. As WACSAW moves through the area in Seg 2, the distributions become more similar even as they are more volatile, and the score decreases. As the calculation moves into the second quiet period, the Wasserstein score again increases as the distributions become more different again. The score crosses the threshold and the peak is again determined. Thus, from this small time series, WACSAW identified 3 segments corresponding to the activity plot in (B).

Fig. 3B illustrates a more complex example of the construction of the WTS, where

the movement transitions from a sleep period to wake and then back to sleep. The data in Fig. 3B is truncated, and values before and after the presented time interval were used in calculating the Wasserstein distance but not plotted. As the algorithm approaches the transition from low variability to high variability, the Wasserstein score (purple curve) increases as the differences in distribution increases. As the time series moves into the middle of the more volatile period, the Wasserstein score decreases as the forward- and backward-looking distributions look more similar. Then as the volatility becomes less in the forward-looking period with the backward-looking period remaining volatile, the Wasserstein score increases again. Lastly the Wasserstein score settles near zero as the volatility becomes very low in both the forward- and backward-looking series after the point of interest moves into the period of low volatility. To define a segment boundary, we determined the peak of the Wasserstein series as the cut point (vertical red lines). Yet, it is clear that all peaks in the WTS are not necessarily meaningful and result in too many segments as well as segments of very small duration. To address this concern, we employed a peak detection algorithm above a threshold which we termed the Change Point Threshold (CPT; horizontal, dashed red line in Fig. 3B). The CPT is defined as the mean of the Wasserstein scores over a 2-day period (see hyperparameter optimization section for a detailed explanation). This threshold is applied over the entire 2-day period and is unique for that individual and for that 2-day period (Fig. 3B), allowing it to vary based on the unique characteristics of both variables. We designate the region of contiguous time points above the CPT to be the change point region, the peak detection algorithm was applied within each change point region (gray regions in Fig. 3B). Thus, within each such region, WACSAW determines and labels the highest Wasserstein score within that region as the change point (vertical red lines).

How sensitive the Wasserstein distance is to differences in distribution can be controlled by the distance order p, lower values of which magnifies differences between distributions. Thus, an appropriately chosen value of p will enable us to magnify the difference between distributions associated with quiescent wakefulness and sleep. In addition, the use of the CPT that are individual specific and the use of the Wasserstein equation can be employed to detect relative differences in volatility in order to identify periods with different activity specific to that individual. In contrast, use of a population-based threshold that signifies sleep periods below it and wakefulness above it would not be adaptive to individual differences.

With this change point detection procedure, we identified regions within which relatively homogeneous volatility is detected, especially in the waking period. However, the segmentation protocol produced too many small regions, with too little data to analyze for classification, because of the highly sensitive nature of the Wasserstein algorithm we employed. Thus, we sought to merge adjacent regions with relatively similar volatility. We used the nonparametric Levene Test for Equal Variance [42] to compare the distributions of first differences, independent of length, between adjacent regions. For segments less than 30 secs, the segment was automatically concatenated to the previous segment. To eliminate influence of extreme outliers on the Levene Test, especially in lengthier segments that increases the occurrence of outliers, the following strategy was employed. We took the absolute value of the first differences of the tilt angle and eliminated the extreme 25%. From the remaining first difference values, we calculated the mean and standard deviation and eliminated data points outside 1 standard deviation from the mean. The Levene Test was run on the resulting distributions from two adjacent segments. If the current segment was not significantly different from the previous segment, at $\alpha = 1e - 25$, two segments were merged. To achieve proper computation of such small values, we used arbitrary precision floating point arithmetic. Though the significance level employed for joining the segments seems extraordinarily strict, it provided the optimal discrimination between segments (see

hyperparameter section for a discussion). If the segment of interest was significantly different from the previous segment, then we compared it with the subsequent segment. If the segments were not different from either the previous or following segments, they were merged. If they were different, they remained their own entity. For example, this process resulted in a reduction of segments from 716 to 118 in a 2-day recording, which was typical of the results we analyzed.

We use Fig 4 to illustrate the consequences of the Levene Test. Fig. 4, In Fig. 4A, each vertical line represents a change point defined by the initial WACSAW algorithm. The vertical dotted red lines in Fig. 4B represent those change points with insignificant difference in variation on either side as determined by the Levene test. Thus, WACSAW removes the segmentation boundary and the distributions of the segments on either side of the boundaries are merged. In Fig. 4C, the four full segments were merged into two segment with the segment boundaries defined by dashed lines eliminated. After the mergers based on the Levene Tests, we are left with a set of change points, and adjacent change points define the boundaries of segments which should contain roughly similar movement behaviour and will ignore brief changes that may disrupt a long period of consistent behavior, such as a rollover event. It is typically robust enough to segment out a brief awakening that has goal-directed behaviors, such as a short visit to the bathroom in the middle of the night. These segments will be subject to sleep-wake classification.

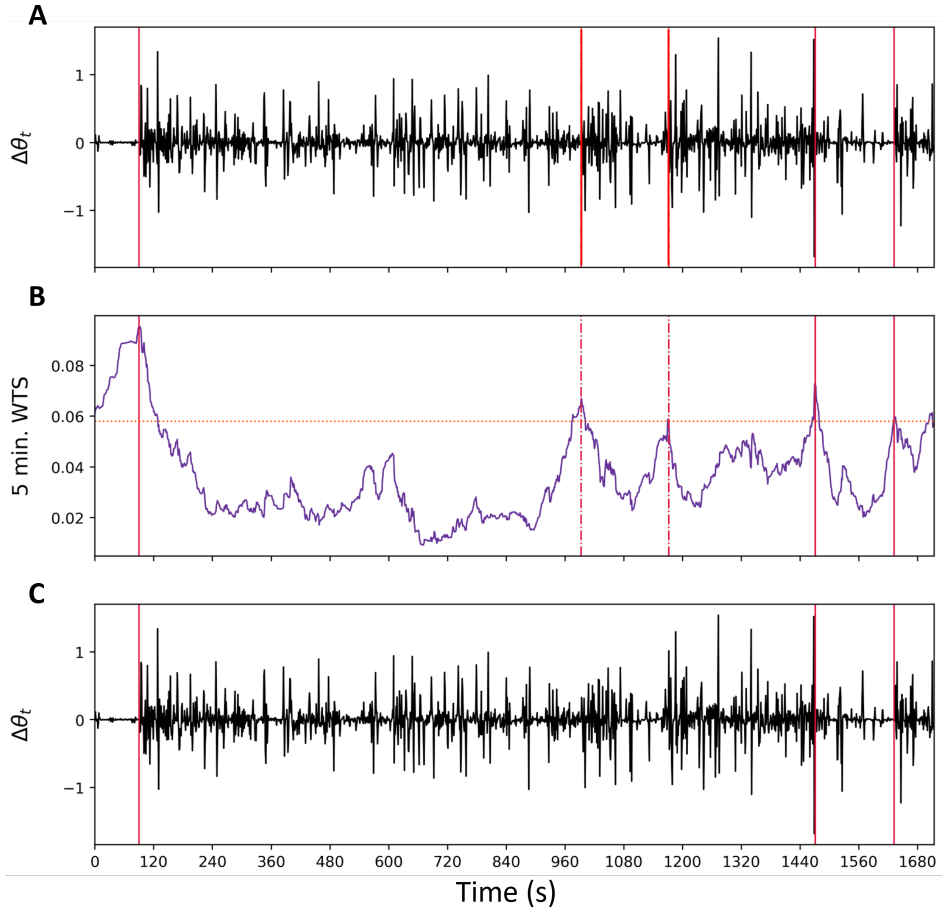


Fig 4. Example of the implementation of the Levene Test combination of segments. (A) To determine whether two segments should be merged, WACSAW starts with the first differences plot for a given segment. The vertical red lines are a reference for where WACSAW assigns segmentation divisions from data in (B). (B) WACSAW converts the first differences data into an optimal transport score time series (WTS = Wasserstein Time Series). Each segment is then compared to the previous and the subsequent segment using the Levene test to determine if segments should be merged. In this example, the horizontal dashed orange line represents the change point threshold. After the Levene test is applied, two of the change points between segments were merged (dashed vertical red lines). (C) The result of the Levene Test in which the final segmentation points identified by WACSAW are illustrated by the solid, vertical red lines.

Segmentation performance

We evaluated the results of the segmentation portion of WACSAW across the six subjects on whom we trained the algorithm. These 6 participants had features that posed challenges to other algorithms. There is no absolute comparator that quantifies the quality of segmentation, though we will show the final results of the algorithm's performance later in the manuscript which depend on proper segmentation. As expected, there are more frequent cuts during daytime wakefulness compared with sleep periods because waking activities vary throughout the day. Segments are of varying lengths which also would be a hypothesized property of daily behaviors. One overnight example

demonstrates the general performance of the segmentation portion of WACSAW (Fig. 5). WACSAW identifies segmentation points at what looks like extended periods of movement volatility, including what appears to be different levels of waking activities from 8-9 AM (Fig. 5A), and another more extended waking period between 3 and 4 AM. Yet, WACSAW does not identify infrequent and brief movements throughout the night as change points. These events signify brief deflections in the first differences of the tilt angle in both the positive and negative direction. These brief movements were not reported in the logs and may reflect roll quick postural shifts such as roll over events. The quick shifts did not elevate the Wasserstein score above the change point threshold, and thus these are not designated as a wakefulness events according to WACSAW.

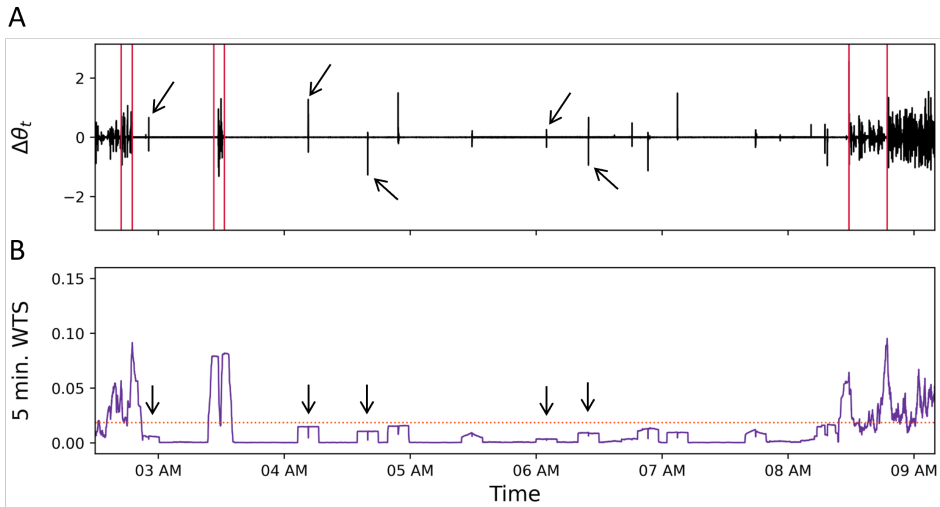


Fig 5. Example of movement that does not result in segmentation. (A) A 7-hour segment of the FD of the tilt angle was chosen to highlight the impact of short duration movements at night. In this trace, there are multiple segments identified by WACSAW with boundaries designated by the vertical red lines. Several brief movements, deflections both up and down in the tilt angle (arrows) are not identified as segments. (B) The Wasserstein score time series from the data in (A). The score does not cross the horizontal threshold line to be designated as a segmentation point (arrows correspond to the arrows in (A)). Thus, the continuity of the segment remains in WACSAW analysis. WACSAW does detect two different movement segments between 8 and 9 AM in this example.

Step 3: Classification

The last step is to classify the segments into sleep and wakeful states. The classification of segments is done in two steps: (A) initial separation of segments into sleep, wake, and undetermined groups and then the (B) final classification based on the sleep and wake groups obtained in step A by using that person's activity data to assign the segments in the undetermined group into sleep or wake categories.

In Step A, we compared each segment with an idealized sleep segment in which the movement values are all zero. For each segment, a thousand subintervals, with a minimum length of 30 seconds within each segment of interest, are randomly selected (Fig. 6A). The Wasserstein distance between the distribution of the middle 95% of the first differences in each subsegment and the idealized sleep distribution (baseline distribution degenerate at zero) was then computed (Fig. 6B). This provides scores reflecting the transport energy required to move each subsegment distribution to the

baseline distribution. We filtered the data by eliminating the extreme 5% of raw first difference values to avoid misclassification of short sleep segments due to the presence of a few large movements. This procedure generates one thousand Wasserstein scores for each segment, and these scores form a distribution reflecting the difference of a segment from that of idealized sleep. For this segment, there are scattered values to the left and right of the peak within each distribution, indicating subsegments with more or less variability than those that make up the prominent peak, but the large peak defines this segment because it is the most frequent value, roughly 0.0075 (Fig. 6C). This distribution is typical of one in which there are short bursts of activity within a larger low-variability segment. Mathematically, the Wasserstein distance used in this step of classification reduces to

$$\mathcal{D}_i^{(\ell)} = \left(\frac{1}{n_i^{(\ell)}} \sum_{t \in S_i^{(\ell)}} |\Delta\theta_t^*|^k \right)^{\frac{1}{k}}, \tag{4}$$

where the value of p is set to k, $\ell = 1, ..., L$ represents the subsegment of interest, $S_i^{(\ell)} \subset \mathcal{S}_i$ represents the time points within a subsegment of $\mathcal{S}_i$, and $n_i^{(\ell)}$ represents the number of points in the subsegment $S_i^{(\ell)}$. Note that $\Delta\theta_t^*$ is the first difference at time t if the first difference is within the middle 95% of data in the segment and zero otherwise. Borrowing from optimal transport phraseology, the value $\mathcal{D}_i^{(\ell)}$ will be referred to as the transport energy associated with the ℓth resampled segment. The collection of $\{\mathcal{D}_i^{(\ell)}\}_{\ell=1}^{L}$ generates an empirical distribution, denoted by $\mathcal{D}_i$, for each segment.

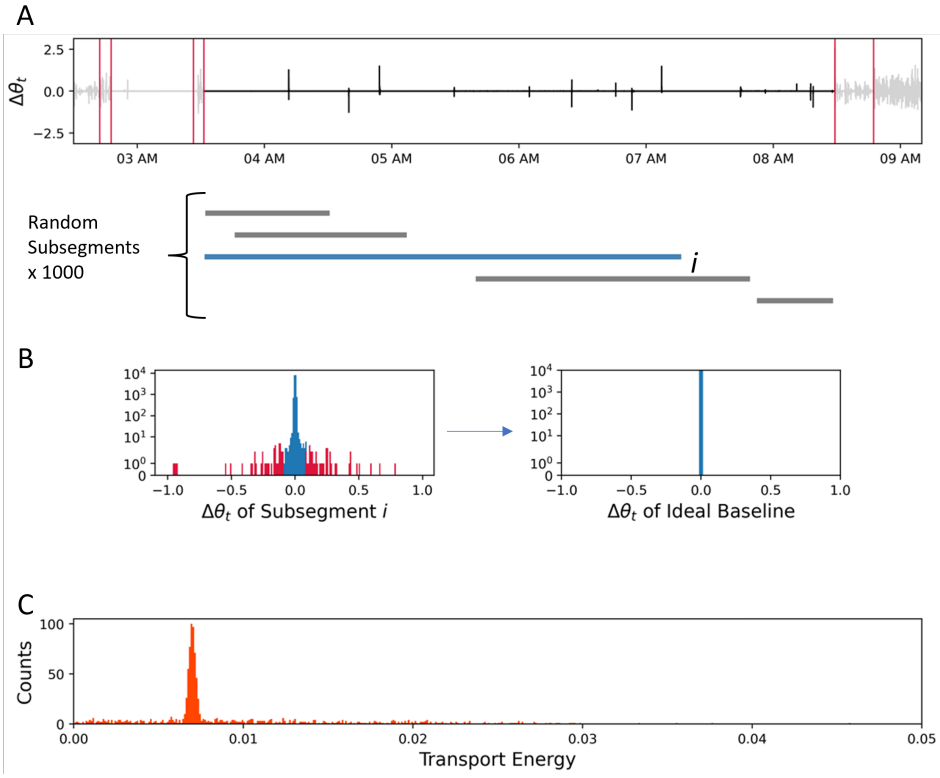


Fig 6. Example of the permutation test to characterize the variability of a segment. (A) The first differences time series for the segment from Fig 5 with the resulting segmentation boundaries after the Levene test step designated with red lines. The segment being analyzed is in black, while grayed out segments will be treated similarly in an independent calculation. Five random hypothetical segments of the 1000 generated by WACSAW are Illustrated under the segment of interest. Each of the randomly generated segments is individually compared to "idealized sleep" by optimal transport methods. (B) The left histogram is the distribution of first difference values from subsegment i where we remove the extreme 5% of points (red bars are censored, blue bars are kept). That distribution is compared to the "idealized sleep" distribution in the right distribution to generate an optimal transport score. (C) The histogram of the optimal transport score for each of the 1000 iterations of subsegment i to characterize the variability of the entire segment, which appears to have a peak around 0.075. There are portions of the segment that have higher and lower variability with scores greater and less than 0.075, but they are a small minority of subsegments. The overall shape of the distribution characterizes the variability of the segment.

Over a full 2-day recording period, WACSAW generates a profile typical of the one presented in Fig. 7A. Histograms from individual segments are differentiated by color. There is a cluster of tall, thin histograms to left side of the figure indicating less difference from idealized sleep. Moving to the right of the distribution profile, the peaks become flatter and broader, indicating more variability throughout the segment. More variability is usually an indicator of some wakeful activity. Note that there are values to the right of 0.05, but the graph has been truncated to highlight the distributions from 0-0.01 of the profile.

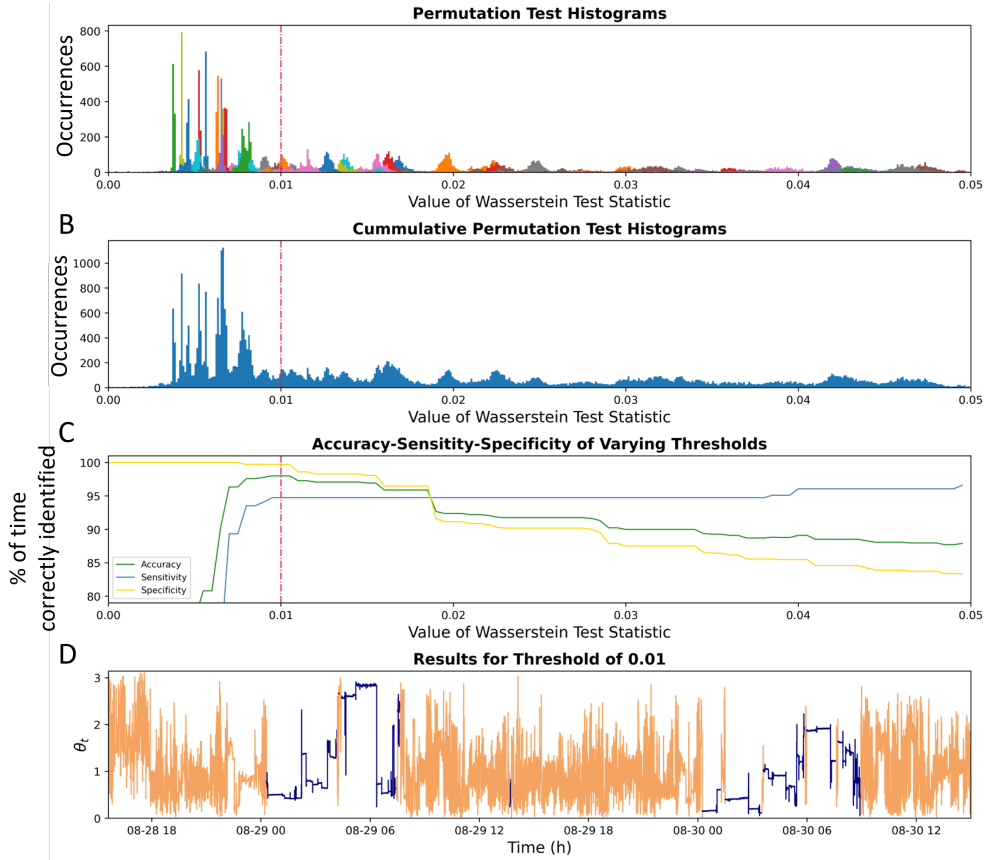


Fig 7. Accuracy resulting from using Wasserstein score cutoff to define classification. (A) Histograms of the Wasserstein scores for segments derived from 2 days' worth of data from Participant 5. To illustrate the relationship between segments, different colors represent subsegments Wasserstein scores from different segments compared to "idealized sleep." Due to limited palette selection, some colors are infrequently duplicated. (B) The histograms for an individual participant for two days' worth of data without identifying individual segments. By eye, there appears to be a difference in the appearance of histograms below a 0.01 score and those above (represented by the vertical red line). (C) We iteratively tested the percent of correct categorization using thresholds of Wasserstein scores from 0 - 0.05 compared against the activity logs. We evaluated overall accuracy (green line), sensitivity to correctly identifying sleep periods (yellow line), and selectivity to correctly identifying wakefulness periods (green line). In this example, the best balance of all three statistics was at a Wasserstein score of 0.01 (vertical red line). (D) Results of using a hard 0.01 threshold to classify sleep and wakefulness overlaid on the tilt angle data gathered for that time series. Orange indicates wakefulness calls and purple indicates sleep periods, as identified by WACSAW. The overall accuracy is ¿90% under these conditions.

Based on observations from Fig. 7B, which shows the the cummulative Wasserstein score histogram, we hypothesized that sleep-like segments were to the left side of the graph while waking segments tended to reside on the right side of the graph. To test this hypothesis, we used a sliding cutoff point to determine if there was a threshold at which the accuracy of sleep and wakefulness calls were optimized using the data from Fig. 7B. The correct designations for each second of the recording were calculated as a percent of total recording time (accuracy), percent of correct sleep calls over total amount of sleep

(specificity), and percent of correct wakefulness calls of the total amount of wakefulness (sensitivity). WACSAW calls were validated using the participant logs or adjusted log as the comparator group for what occurred during that time period. As the Wasserstein value went from 0 up through 0.01, the accuracy of each of the measures improved, except for sleep sensitivity because when the program calls everything sleep it will always get the sleep segments correct. The optimal value appeared to be 0.01 for this individual, which resulted in the calls in Fig. 7C. For this individual dataset, it resulted in a >95% accuracy across the 2 days. Yet, this hard cutoff did not work for all individuals. We observed several shallow drops in accuracy between 0.01 and 0.02 (Fig. 7C) indicating an overlap of sleep and wakeful segments because there was not a threshold that optimized the sleep and wakefulness calls for every individual. For the example participant, setting the threshold to a transport energy to 0.01 resulted in a >97% accuracy (Fig. 7D). But in other participants, a threshold of 0.01 reduced the accuracy, suggesting that an absolute population-based threshold was impractical. However, two cut-off values can be identified based on 2-day recoding periods for all individuals in the training data set, with one value (0.01) almost always containing sleep segments to the left of it, and the other value (0.02) almost always containing wake segments to the right of it. These two cut-off values were employed to group all segments into three categories: (1) Sleep, (2) undetermined, and (3) wake.

These designations will be used in later processes as starting points for the classification process. In the second stage of the classification process, we employed a process that transformed the movement distribution from each segment into a single curve, by obtaining the modulus of the characteristic function of that distribution. For a given segment Si, the characteristic function as of the distribution of first difference ϕ_i of the tilt angle is computed using the formula

$$\varphi_i(\xi) = \frac{1}{n_i} \sum_{t \in S_i} e^{-j \cdot \xi \cdot \Delta\theta_t}, \tag{5}$$

where $\mathcal{S}_i$ is the ith segment, n_i is the number of observations in $\mathcal{S}_i$, j is the imaginary unit $\sqrt{-1}$, and ξ a real number, with $\Delta\theta_t$ defined as before. Since the domain of the characteristic function is the same for all segments, irrespective of length, the use of the characteristic functions facilitates comparison and classification across all segments regardless of size. For example, a 7 hr segment of one time series recording (Fig. 8A), there are examples of highly variable segments (orange, Fig. 8B), subtly variable segments (green, Fig. 8C), and low variable segments (blue, Fig. 8D) and activities were validated by participant log. The distribution data are transformed by Eq. 5 that resulted in the modulus of characteristic functions, $|\varphi_i|$, and plotted (Fig. 8E-G).

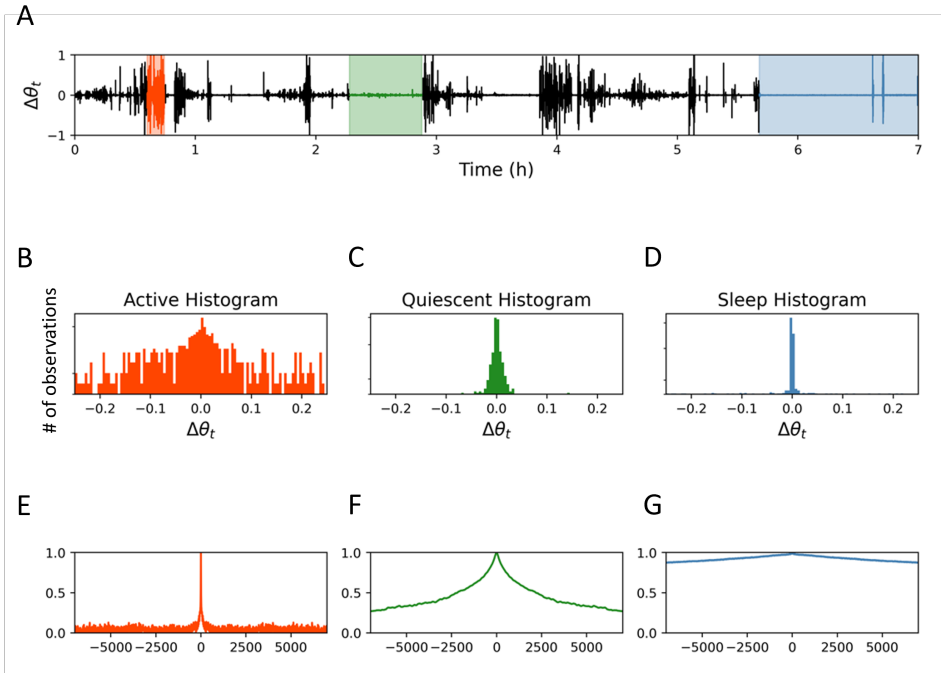


Fig 8. Example of the conversion of first differences time series of the tilt angle into characteristic functions. (A) A 7-hour segment of first differences data with examples of high variability, moderate variability, and low variability segments as noted in the participant's log. (B-D) Histograms of the distributions from the segments of the corresponding activity levels. The color of the histogram aligns with the background color behind the subsegment of data graphed. (E-F) Each distribution is transformed (squiggly line) into a two dimensional representation, referred to as a characteristic function, which can be compared to one another in their shape and values at respective points along the graph.

The modulus of the characteristic function (MCF) obtained from a given segment can be thought of as a curve that captures the movement patterns within that segment. Thus, the MCF curves from segments that are most likely sleep should cluster together and those from segments that are most likely from wakeful periods should cluster together as well. However, there are segments that will lie in between. To classify such segments, the following approach was taken. The modulus of the characteristic functions derived from segments with transport energy D_i averages less than 0.01 (likely sleep; Fig. 7A) are averaged to nucleate a cluster and modulus characteristic functions derived from segments with D_i from 0.02-0.1 (likely wakefulness; Fig. 7A) are averaged to nucleate the second cluster. These two nucleated functions (or curves) are taken as clustering objects and assumed to be ground truth labels from detailed sleep logs. Characteristic functions for other segments were given the same label as the closest nucleated function. Note that while the 0.01 threshold was determined universally, the nucleated clusters for sleep and wake are driven mostly by an individual's movement patterns. Thus, the clustering is to a great extent individually tailored.

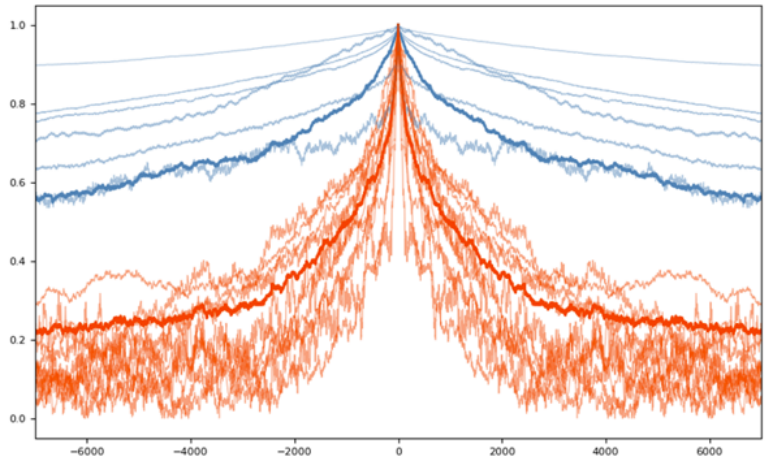


Fig 9. An example of the clustering of characteristic functions to classify sleep or wakefulness. Segments defined by WACSAW are classified as sleep or wakefulness based on the clustering of characteristic functions. Each line is a different segment. The bold blue and orange lines are the averages that nucleate the binary clusters. The average characteristic functions of segments with a Wasserstein score below 0.01 is calculated to nucleate the sleep cluster (thick blue line). The average characteristic function of segments with a Wasserstein score from 0.02-0.06 nucleates the wakefulness cluster (thick orange line). The characteristic function from each segment is then clustered using these averages as the anchor. Segments that cluster with the wakefulness characteristic function are classified as wake (lighter orange lines) and those that cluster with the sleep characteristic function are labeled sleep (lighter blue lines).

Results

WACSAW performance from activity data

In the above sections, we have described how WACSAW was developed, the steps it uses to distinguish sleep and wakefulness from activity data, and provided some examples of how the process operates. Now, we present how WACSAW performs across multiple individuals across different experiments. We used activity data from 6 individuals to develop WACSAW. We monitored the activity across 48 continuous hours. These 6 individuals were designated the "Development set" and were chosen because their activity recordings and logs presented known situations that have challenged past actigraphy algorithms when classifying sleep and wakefulness. Such circumstances include, participants involved in extended activities of quiet wakefulness, split sleep at night, sleep in older individuals, sleep in females, and individuals with some difficulties sleeping at night (though this last one did not rise to the level of a sleep disorder diagnosis), among other traits. Table 1 lists the participants and their demographics.

WACAW development	Participant id	% correct Accuracy (Raw)	% correct Sensitivity (raw)	% correct Specificity (raw)	% correct Accuracy (Adjusted)	% correct Sensitivity (Adj)	% correct Selectivity (adj)
Development Cohort	1	96.9	94.6	98.2	95.8	94.3	96.7
	2	96.7	95.9	97.0	98.8	98.7	98.9
	3	99.1	98.4	99.5	99.5	99.5	99.5
	4	96.0	93.3	97.5	97.6	96.8	98.0
	5	95.5	96.5	95.1	95.8	97.2	95.3
	6	96.5	99.0	95.5	98.4	99.0	98.1
	Mean	96.8	96.3	97.1	97.7	97.6	97.7
	Median	96.6	96.2	97.2	98.0	97.9	98.1
	Std Dev	1.2	2.2	1.7	1.5	1.9	1.5
	MAD	0.8	1.7	1.3	1.2	1.5	1.1
	min	95.5	93.3	95.1	95.8	94.3	95.3
	max	99.1	99.0	99.5	99.5	99.5	99.5
	% > 95%	100.0	100.0	100.0	100.0	100.0	100.0
Development Cohort Validation	1	92.4	94.7	91.3	94.4	97.3	93.0
	2	97.2	93.5	98.8	98.1	95.3	99.2
	3	96.6	99.4	95.3	96.4	99.6	95.0
	4	94.4	91.2	96.4	96.8	99.0	95.7
	5	90.6	97.8	87.6	98.7	98.3	98.9
	Mean	94.2	95.3	93.9	96.9	97.9	96.4
	Median	94.4	94.7	95.3	96.8	98.3	95.7
	Std Dev	2.8	3.3	4.4	1.6	1.7	2.7
	MAD	2.1	2.5	3.2	1.2	1.2	2.0
	min	90.6	91.2	87.6	94.4	95.3	93.0
	max	97.2	99.4	98.8	98.7	99.6	99.2
	% > 95%	83.3	83.3	83.3	83.3	83.3	83.3

Std Dev = Standard deviation
MAD = Mean absolute deviation from the median

Table 2. WACSAW results from the development cohort.

Within this Development set, WACSAW performed admirably (Table 2). WACSAW classifies sleep or wakefulness on a second-by-second basis and produces results without human intervention. As mentioned above, we have compared the output to both the raw log entries from the participants(raw) as well as to an investigator adjusted log that was defined by a set of rules (see above, adj). We present both sets of data for comparison and note that the adjusted data set results in only a marginal improvement. The overall accuracy of WACSAW is 96.8% correct classification with the raw data and 97.7% in the adjusted data. The correct classification of sleep episodes (specificity) is above 96% across this set and the correct classification of wakefulness (selectivity) is above 97%. The median reflects that there are unlikely to be single values that are responsible for pulling the mean in one direction or the other. In addition, the standard deviation (Std dev) and median absolute deviation (MAD), which is the average median of the absolute deviations from median accross the group, suggest that the results are accurate consistently across individuals. Given the age and situational differences, it suggests that WACSAW can somewhat adapt to individual circumstances. The minimum and maximum accuracies for this cohort are all above or near 95%, which suggests consistency within the group that WACSAW was trained on. A high percentage of the values within the cohort are above 95%, suggesting very good classification. Given that the raw values and adjusted values are very similar and that log recordings are subject to errors because participants may not or are unable to document specific times for changes in activity, we will only discuss the adjusted values, but the raw data will be presented for the performance of WACSAW.

Five out of the 6 participants within the Development cohort had a second data collection period that was independent of the original data collection period on which WACSAW was developed. As an initial test of WACSAW's performance, we applied WACSAW to this independent data collection period in the participants with the same

properties as the Development cohort. Overall accuracy, specificity, and selectivity were all above the 95% accuracy threshold for the entire group with only 1 participant records below this threshold for the 48 hour recording period. The standard deviation (SD) and MAD remain low across this small population. This test shows that the performance of WACSAW is not totally reliant upon the training data and can perform well on independent data, albeit obtained from the same subjects used in the development cohort.

To test whether WACSAW can perform equally well on independent data from individuals without the same underlying properties, we applied WACSAW to 48-hour recordings from 16 independent individuals (Table 3). Across the "Independent validation" data set, WACSAW maintained its average of greater than 95% accuracy across overall performance, sensitivity, and selectivity. The std dev and MAD remained low. There is one individual, participant 11, that had one day in which WACSAW demonstrated a significant underperformance. In a posthoc interview, it was determined that the person was sick over the data collection period. It highlights that WACSAW has room for improvement under specific circumstances. Under these circumstances, the interim metrics, like the cumulative histogram of transport energy discussed in Fig. 7, may suggest that results may not be as reliable and are worth further investigator validation (see Fig. 11B below). Despite the 24-hr underperformance, WACSAW performance in all individuals were above 86% with an overall average over 95%.

	Participant id	% correct Accuracy (raw)	% correct Sensitivity (raw)	% correct Specificity (raw)	% correct Accuracy (adj)	% correct Sensitivity (adj)	% correct Selectivity (adj)
Independent Validation	7	97.7	94.2	99.4	99.4	99.2	99.5
	8	96.6	94.4	97.9	97.7	97.3	98.0
	9	97.4	94.4	99.0	98.8	97.1	99.6
	10	97.3	94.8	98.2	97.2	94.7	98.1
	11	85.3	68.4	96.0	86.3	69.4	97.0
	12	98.9	98.3	99.2	98.9	98.4	99.1
	13	96.6	98.6	95.4	98.4	99.5	97.7
	14	95.5	99.6	94.3	96.7	99.9	95.7
	15	95.5	88.1	100.0	98.6	97.2	99.3
	16	86.4	98.3	82.4	98.8	97.2	99.7
	17	98.8	97.9	99.2	98.7	98.0	99.2
	18	97.6	98.8	96.9	98.8	99.3	98.5
	19	99.4	99.1	99.5	99.1	97.7	99.7
	20	98.9	96.8	100.0	98.9	97.0	100.0
	21	94.8	96.7	94.1	97.0	98.5	96.3
	22	86.4	91.8	84.0	89.9	95.5	87.3
	Mean	95.2	94.4	96.0	97.1	96.0	97.8
	Median	97.0	96.7	98.1	98.7	97.5	98.8
	Std Dev	4.7	7.6	5.4	3.7	7.2	3.1
	MAD	3.1	4.0	3.4	1.8	2.8	1.7
	min	85.3	68.4	82.4	86.3	69.4	87.3
	max	99.4	99.6	100.0	99.4	99.9	100.0
	% > 95%	75.0	56.3	75.0	87.5	87.5	93.8

Std Dev = Standard deviation
MAD = Mean absolute deviation from the median

Table 3. WACSAW performance in the independent validation cohort.

It is possible that we have been evaluating a set of data for which is it simple and easily definable to identify sleep and wakefulness. Therefore, we compared WACSAW's output to the sleep and wakefulness designations produced by the clinically validated Actiwatch that is produced and maintained by Respironics. The Actiwatch is a popular, validated actimeter which uses piezoelectric accelerometers to obtain movement counts which are then used to determine sleep and wakefulness. For a detailed treatment of the Actiwatch algorithm, we refer to you to the Actiwatch documentation [35]. The data

from these samples was analyzed in earlier charts but had not been benchmarked to an activity device that has been validated and approved for use in a clinical and research setting. For this period of data collection, participants were wearing both watches at the same time on the same wrist. The data from the Actiwatch was processed through the company's analysis software to obtain the classifications whereas WACSAW was applied to the GENEActiv MEMS data as described above.

We maintained the same divisions in the population because the first 5 participants were the majority of the data on which WACSAW was originally developed. WACSAW performs well on the data on which it was optimized as discussed above (Table 4), and it outperforms the Actigraph by roughly 10% on measures of accuracy, sensitivity and selectivity. Across the population, WACSAW performs better between individuals as the Std dev and MAD are much lower with WACSAW designations than with Actiwatch classifications. Nearly all the values for the WACSAW designations are greater than 95% accuracy in each of the criteria, whereas a minority, if any, of the Actiwatch cross that threshold.

		WACSAW			Actiwatch		
WACSAW vs Actiwatch	Participant id	% correct Accuracy (adj)	% correct Sensitivity (adj)	% correct Specificity (adj)	% correct Accuracy (adj)	% correct Sensitivity (adj)	% correct Selectivity (adj)
Development cohort	1	95.8	94.3	96.7	87.8	79.6	92.4
	2	98.8	98.7	98.9	92.2	92.4	92.0
	3	99.5	99.5	99.5	88.9	96.0	85.4
	4	97.6	96.8	98.0	85.2	86.4	84.6
	5	95.8	97.2	95.2	73.8	89.7	67.2
	Mean	97.5	97.3	97.6	85.5	88.8	84.3
	Median	97.6	97.2	98.0	87.8	89.7	85.4
	Std Dev	1.7	2.0	1.7	7.0	6.2	10.2
	MAD	1.3	1.4	1.3	4.4	4.5	6.5
	min	95.8	94.3	95.2	73.8	79.6	67.2
	max	99.5	99.5	99.5	92.2	96.0	92.4
	% > 95%	100.0	80.0	100.0	0.0	20.0	0.0
Independent validation cohort	7	99.4	99.2	99.5	92.2	94.0	91.4
	9	98.8	97.1	99.6	93.9	87.6	97.5
	10	97.2	94.7	98.1	80.6	84.3	79.2
	16	98.8	97.2	99.7	89.5	88.9	89.9
	17	98.7	98.0	99.2	89.5	88.0	90.3
	18	98.8	99.3	98.5	87.6	91.6	85.3
	19	99.1	97.7	99.7	91.9	91.4	92.0
	20	98.9	97.0	100.0	89.0	72.6	98.7
	21	97.0	98.5	96.3	90.2	80.5	94.4
	22	89.9	95.5	87.3	73.4	84.2	68.2
	Mean	97.7	97.4	97.8	87.8	86.3	88.7
	Median	98.8	97.5	99.4	89.5	87.8	90.8
	Std Dev	2.8	1.5	3.9	6.2	6.3	9.2
	MAD	1.3	1.1	1.9	3.8	4.5	6.1
	min	89.9	94.7	87.3	73.4	72.6	68.2
	max	99.4	99.3	100.0	93.9	94.0	98.7
	% > 95%	90.0	90.0	90.0	0.0	0.0	20.0
Full cohort	Mean	97.6	97.4	97.7	87.0	87.1	87.2
	Median	98.8	97.2	98.9	89.0	88.0	90.3
	Std Dev	2.4	1.6	3.2	6.3	6.1	9.4
	MAD	1.4	1.2	1.8	4.1	4.6	6.6
	min	89.9	94.3	87.3	73.4	72.6	67.2
	max	99.5	99.5	100.0	93.9	96.0	98.7
	% > 95%	93.3	86.7	93.3	0.0	6.7	13.3

Table 4. Comparison of WACSAW performance to Actiwatch designations.

When we evaluated 10 individuals from the independent validation cohort, the pattern described above was maintained. The overall performance of WACSAW

remained greater than 95% accuracy with low variance and high proportion of the individuals clearing the 95% threshold. This dataset includes the individual for which WACSAW had difficulty identifying sleep according to comparisons with participant recorded log data. Even so, WACSAW outperformed the Actiwatch analysis of the movements over the same time period. The overall performance of the Actiwatch algorithm was lower and the variation within the population is higher with almost none of the values crossing the 95% accuracy threshold. When the data are combined in the "Full cohort," our interpretation is that the conclusions remain the same. Thus, WACSAW is not analyzing data that is easy to categorize as the Actiwatch found them to be easy to categorize and thus produce higher accuracies.

We hypothesize that one big difference between the accuracy of WACSAW and the accuracy of the Actiwatch designations appeared to be in classifying quiescent wakefulness. Fig. 10 presents 3 instances of large differences between WACSAW and Actigraph. Each of the examples shows the raw MEMS movement recording, how the participant categorized behavior at the time, and how the Actiwatch system classified the movement data. The examples include one participant reading in bed (Fig. 10A), and two other instances where the Actiwatch classification differed from the participant log during activities such as watching TV (Fig. 10B) and desk work (Fig. 10C), but WACSAW correctly classified them.

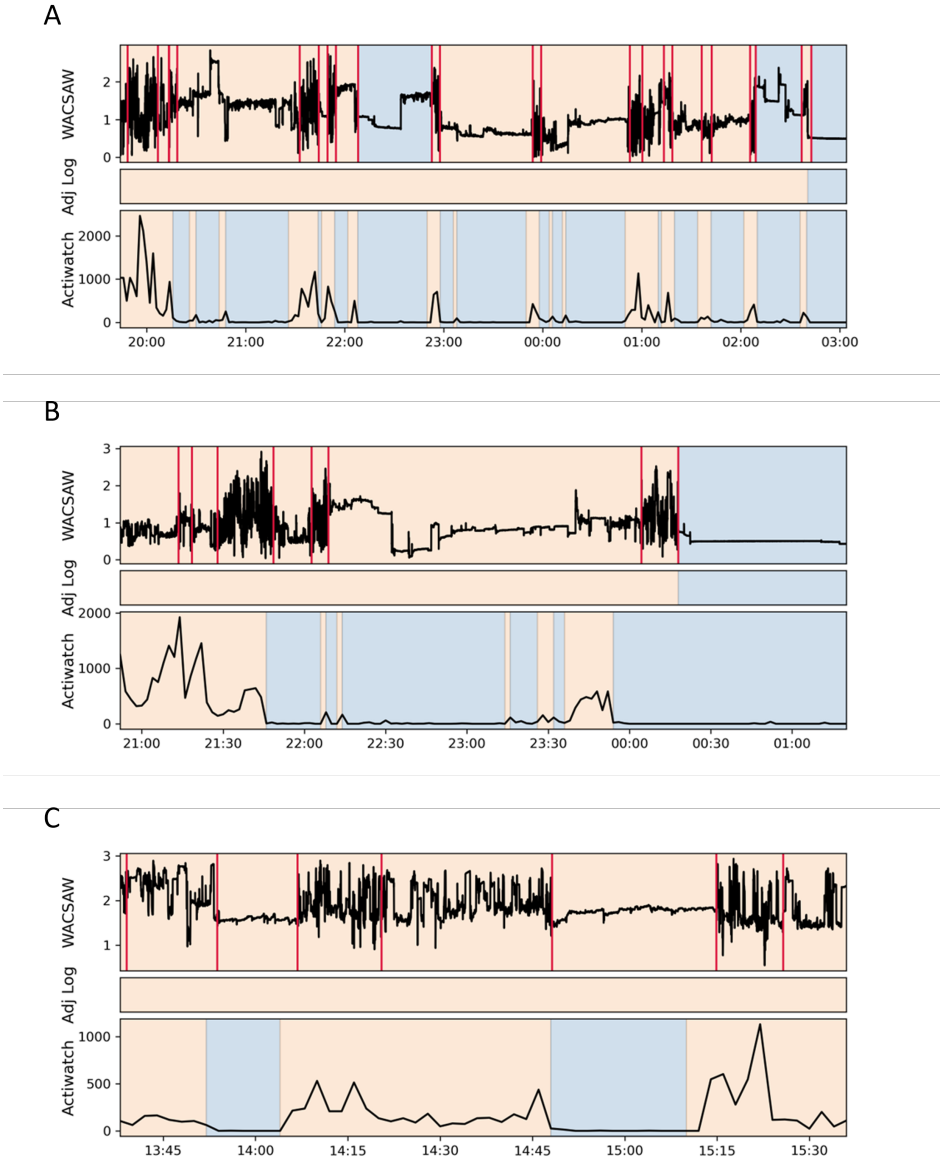


Fig 10. Comparison of behavioral categorization between WACSAW and Actiwatch. We chose 3 raw data examples (A-C) to illustrate differences in the performance of WACSAW compared to Actiwatch with a particular focus on the challenges of quiet wakefulness. Orange background indicates a wakefulness categorization by WACSAW or the Actigraph algorithm or that the person logged they were awake (middle bar in each panel. A blue background indicates sleep for all three. In (A), a participant was reading in bed during this period. In (B), a second participant was watching TV in their couch during the wakefulness period leading up to sleep. While in (C), a third participant was working at their desk during the entire recording period shown. Vertical red lines in the WACSAW trace indicate segmentation breaks determined by the algorithm. The Actigraph assigns many more sections as sleep when the participant was engaged in quiet waking activities (see middle bar for actual activity).

Lastly, we compared our outcome to an existing analysis program that also accepts

data from the GENEActiv watch to determine if the frequency of data collection was the only difference in the accuracy of the predictions. We submitted the data used for earlier WACSAW predictions to the GGIR program and compared its output to the participant logs using the adjusted measurements. We found that WACSAW was approximately 5% better in overall accuracy, roughly 12% more accurate with sleep, and approximately 2% better with selectivity, which identifies wakefulness. We compared the visual outputs from each program (Fig. 11). GGIR makes substantially more state changes, especially during the night compared to WACSAW. For instance, in participant 19, the first night of sleep is segmented quite frequently using GGIR compared to how WACSAW determines sleep and wakefulness. It may be in these numerous transitions that leads to WACSAW's distinction from GGIR.

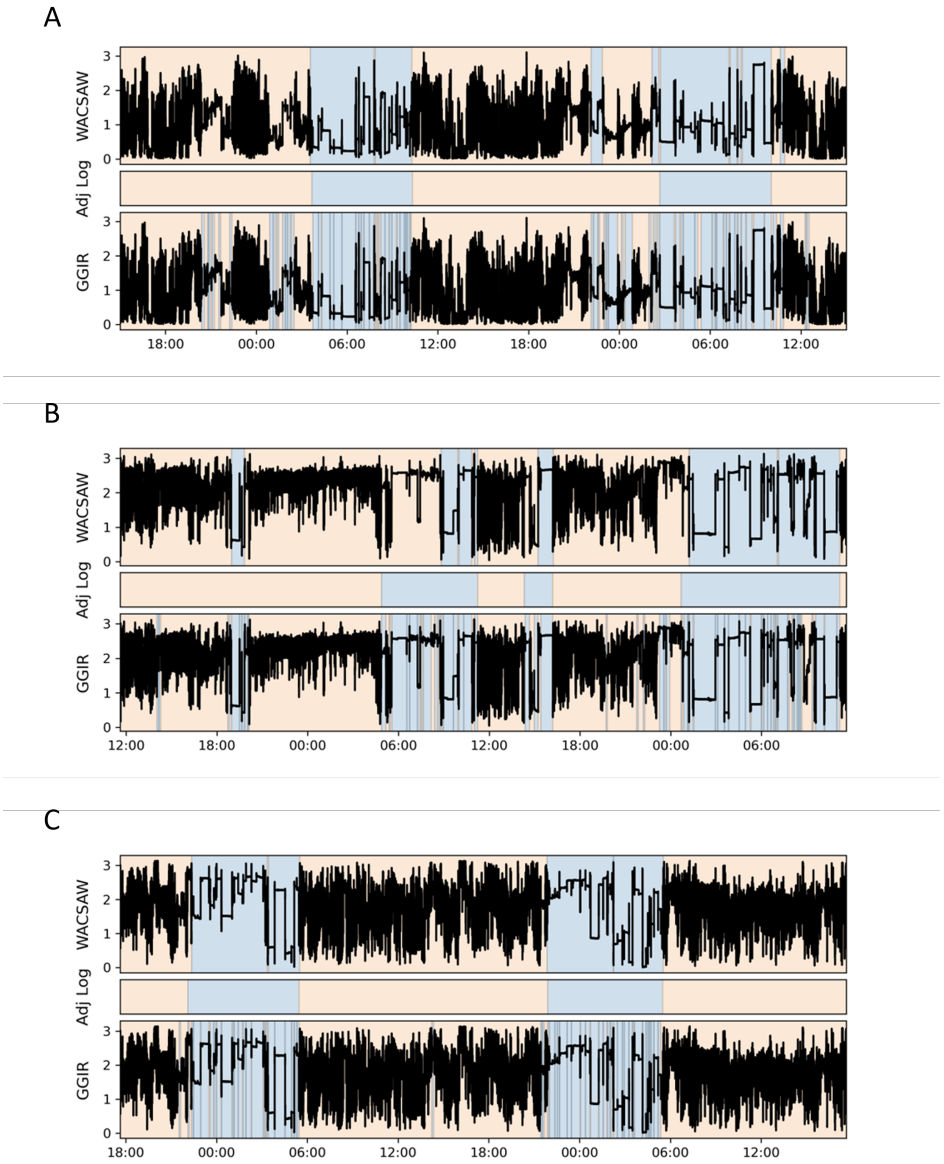


Fig 11. Comparison of time series of sleep/wakefulness output from WACSAW and GGIR. We directly compared the output from 3 participants, Part 5, which was an older individual, Part 11, in which WACSAW demonstrated its worst performance, and Part 19, in which WACSAW did very well. An orange part of the trace indicates that the given software called the section wakefulness and a blue portion of the trace indicates that the software labeled the section as sleep.

It needs to be highlighted at this point that our validation criteria are participant logs. Because it is based on participant recollection and documentation, the logs may reflect longer stable periods of behavior rather than the numerous unrecognized or unrecollected sleep and wakefulness transitions. There is no indication in the logs that there is this number of transitions, but logs are not the definitive determination of state changes. What is apparent from the comparison between GGIR and WACSAW is that they have different tolerances for activity transitions, which may be delivering different types of information to the user. As discussed later, WACSAW needs to be validated

using polysomnography (PSG) to determine if the brief movements are indicative of wakefulness. It will also be interesting to determine if those brief wakefulness periods are meaningful to other health or cognitive parameters.

		WACSAW			GGIR		
WACSAW vs GGIR	Participant id	% correct Accuracy (adj)	% correct Sensitivity (adj)	% correct Specificity (adj)	% correct Accuracy (adj)	% correct Sensitivity (adj)	% correct Selectivity (adj)
Development cohort	1	95.8	94.3	96.7	88.0	78.4	92.8
	2	98.8	98.7	98.9	96.7	91.0	99.1
	3	99.5	99.5	99.5	95.9	94.2	96.7
	4	97.6	96.8	98.0	92.0	90.1	93.0
	5	95.8	97.2	95.3	86.6	90.2	85.0
	6	98.4	99.0	98.1	89.4	77.4	95.1
	Mean	97.7	97.6	97.7	91.4	86.9	93.6
	Median	98.0	97.9	98.1	90.7	90.2	94.0
	Std Dev	1.7	2.0	1.7	4.5	6.0	5.3
	MAD	1.4	1.6	1.3	3.9	3.4	3.8
	min	95.8	94.3	95.3	86.6	78.4	85.0
	max	99.5	99.5	99.5	96.7	94.2	99.1
	% > 95%	100.0	80.0	100.0	40.0	0.0	40.0
Indepentent Valication cohort	7	99.4	99.2	99.5	96.4	91.3	98.9
	8	97.7	97.3	98.0	91.2	84.6	95.1
	9	98.8	97.1	99.6	92.9	81.7	99.9
	10	97.2	94.7	98.1	92.4	85.2	94.9
	11	86.3	69.4	97.0	88.6	83.9	91.7
	12	98.9	98.4	99.1	91.0	87.6	92.8
	13	98.4	99.5	97.7	88.4	82.1	92.2
	14	96.7	99.9	95.7	91.5	93.6	90.8
	15	98.6	97.2	99.3	88.9	70.0	98.4
	16	98.8	97.2	99.7	92.7	85.6	97.3
	17	98.7	98.0	99.2	94.6	88.3	98.1
	18	98.8	99.3	98.5	91.9	88.9	93.5
	19	99.1	97.7	99.7	96.9	93.2	98.6
	20	98.9	97.0	100.0	89.5	70.1	100.0
	21	97.0	98.5	96.3	92.6	81.0	97.4
	22	89.9	95.5	87.3	85.7	86.4	85.4
	Mean	97.1	96.0	97.8	91.6	84.6	95.3
	Median	98.7	97.5	98.8	91.7	85.4	96.2
	Std Dev	3.7	7.2	3.1	3.0	6.8	4.0
	MAD	1.8	2.8	1.7	2.2	4.8	3.3
	min	86.3	69.4	87.3	85.7	70.0	85.4
	max	99.4	99.9	100.0	96.9	93.6	100.0
	% > 95%	87.5	87.5	93.8	12.5	0.0	56.3
Full cohort	Mean	97.2	96.4	97.8	91.5	85.2	94.8
	Median	98.5	97.5	98.3	91.7	86.0	95.1
	Std Dev	3.1	6.1	2.6	3.2	6.6	4.1
	MAD	1.7	2.4	1.6	2.6	5.2	3.3
	min	86.3	69.4	87.3	85.7	70.0	85.0
	max	99.5	99.9	100.0	96.9	94.2	100.0
	% > 95%	90.9	86.4	95.5	18.2	0.0	54.5

Table 5. Comparison of WACSAW performance to GGIR calls.

Discussion

WACSAW is an adaptive sleep-wake classification algorithm with novel applied statistical approaches to segmenting and classifying movement data into sleep and wakefulness. WACSAW delivered an average of >96% accuracy, sensitivity, and selectivity across the entire healthy person cohort when compared to detailed activity logs. Moreover, WACSAW reduced individual error, which resulted in a smaller standard deviation across the entire population. When compared to the clinically

approved Actiwatch, WACSAW had a higher sleep/wake state accuracy and a lower standard deviation between individuals within our participant population, indicating better and more personalized performance in categorizing sleep and wakefulness.

We developed WACSAW from data collected using the GENEActiv actimeter, which uses the MEMS to detect changes in movement at a subsecond sampling rate. The sampling property has distinct advantages in classifying sleep and wakefulness states compared to traditional piezoelectric methods. First, we could compose a more detailed representation of movement variability within a smaller time window, which allows for a faster determination of segmentation boundaries between two periods of different activity levels. Second, we could generate probability distributions within smaller time windows. Thirdly, higher frequency collection provides robust distributions of movement statistics in a short time window, making the Wasserstein distance more accurate in determining changes in underlying distributions. Lastly, segmentation becomes more localized, which minimizes sleep-wake overlap in any given region and provides a more accurate partitioning of sleep and wakeful regions. MEMS technology has permitted us to explore statistically based approaches that accommodates individual differences rather than regression-based approaches. Because WACSAW is agnostic to the exact nature of the data, it could also be applied to other classification systems that rely on variability changes.

Another benefit of the WACSAW process is that it produces what we are calling "interim metrics." Interim metrics are outputs that are generated as WACSAW is determining the definition and classification of each of the segments. Examples of these metrics include the Wasserstein time series, the transport energy histograms, and the clustering of the characteristic functions. These interim metrics are unique to the initial data analyzed and could be used to further characterize the sleep and activity over a given period of time or potentially as a measure of confidence for the designations assigned by WACSAW. For example, if the transport energy histograms are overlapping or very closely spaced at a given Wasserstein score, it may indicate poorer sleep/wakefulness segmentation boundaries that introduces errors in interpretation. In such a case, the investigator could dedicate their time and effort to this data set and feel confident that other data sets are categorized well. Another example would be when the shape and overlap of the characteristic functions may indicate poor classification of sleep and wakefulness. In fact, statistical techniques could be derived or applied that would characterize or represent the movement data from the interim metrics. These representations could be a further objective metric to compare with an outcome such as sleep quality or another subjective measure. Another interesting interim metric is the Wasserstein time series (WTS), which is the time series generated by calculating the optimal transport score across time. Though summing the first difference of the tilt angle would give us an indication of how much movement occurred during the night, the WTS gives an indication of the relative magnitude of the movement and how long it persisted. Additionally, the Wasserstein series has different shapes for different types of movements. For example, brief movements are more box shaped while longer duration movements are more rounded and have an "m" shape (see Fig. 5C). These values can then be used in subsequent calculations or time series analyses. We believe that these metrics are a strength of the WACSAW approach and offer an opportunity for more sophisticated and meaningful analyses.

Based on the analyses above, WACSAW offers a different approach and potentially has advantages over current methods, such as regression-based algorithms and neural network approaches. Regression-based approaches worked well for the lower sampling rates of piezoelectric devices which have a strong record for gathering information. Some of the most utilized algorithms are subject to flaws due to the fact that they are regression-based, such as the Sadeh [22] and Cole-Kripke [23]. In fact, a comparison of

popular sleep-wake algorithms concluded that for healthy individuals, the Sadeh algorithm still remains optimal [29]. Regression tries to find one equation that best fits the population, which can unknowingly lead to large errors in some individuals that do not conform to the average. The algorithms do have an acknowledged weakness in that sleep sensitivity ranged from 0.86 and 0.99 while specificity performed more poorly and ranged widely between 0.27 and 0.68, likely due to a misclassification of quiet wakefulness, for example desk work, watching television, and reading [24, 28].

Additionally, machine learning and deep learning techniques have been applied to actigraphy data. The most comparable attempts have used a single data modality or single channel data [26, 43, 44]. In summary, these papers use different sized training populations, different data sources, and applied numerous different approaches to analyze the data. The results demonstrate that there are analyses amongst the methodologies that show improvement over regression techniques. Only some of the analysis methods showed the improvements without a consensus between the approaches and techniques. Another consideration is that it is difficult to determine where the errors in approaches are. WACSAW demonstrates better numbers with a couple of caveats, WACSAW has not been validated against polysomnography (PSG), which could reduce the overall performance because small sections of the record would show differences. For example, a participant could have woken for a short period of time without movement and not recorded that in their logs, which may bring down WACSAW's overall performance to equivalent levels. On the other hand, WACSAW's approach allows those types of problems to be identified and purposeful changes to the approach to address such issues. In comparison to neural network-based algorithms, WACSAW has many fewer parameters that are also more explainable and that can inform researchers and users about the biology of the system.

WACSAW has aspects within the algorithm that give it properties of having individualized sleep and wakefulness designations. The threshold for designating segment divisions is based on the activity profiles for the two-day span for that individual. Moreover, there are additional opportunities to further adapt WACSAW to individuals or specialized situations, such as sleep disorders, using hyperparameters. For example, increasing the order (p) in the Wasserstein equation 3 would make it more sensitive to smaller changes in movement distribution. These changes could be used to adapt WACSAW to improve performance in such circumstances, such as the fragmentation that can take place with sleep in aging or sleep disorders. In addition, the program has the potential to be adapted to evaluate different properties depending on the user's interests. For example, expanding the window of the Wasserstein time series around the point of interest would widen the area to detect a change between segments and allow the program to focus more on larger time scales, such as those for sleep time and rise time. This change would ignore short-term changes in distribution to preferentially identify larger changes, which would result in fewer segments and support identifying longer overall sleep periods rather than brief awakenings, short naps, or nighttime restroom visits. The steps taken to adapt WACSAW to the individual reduced the variability of accuracy within the population below that seen in the Actiwatch analysis despite the higher accuracy. Thus, WACSAW can be further adapted to specific research or clinical goals.

The approach presented here has limitations, and future investigations may validate or mitigate its generalizability. The most obvious limit is the validation against PSG. Our validation using logs is a practical and opportunistic method that both allowed us to progress on WACSAW but is also a good validation method for many of the characteristics that researchers and participants care about. The weakness is that it does not have the moment-to-moment validation of brain state that can reveal short or brief periods of wakefulness and define sleep latency, which is an important parameter

for specific sleep disorders. In fact, we adjusted sleep and wakefulness periods based on a 20 minute window with the acknowledgement that sleep logs can be inaccurate. That said, it only altered accuracy percentages by less than 2%. WACSAW has demonstrated the potential to identify nighttime brief awakenings and potential restlessness in individuals. However, it has thus far been validated to detailed activity logs, thus nighttime states have not been validated against PSG.

Another limitation is the sample size of participants. We were cognizant to include people of different ages, sexes and behavior types. We are aware of the limitations of other algorithms and wanted to specifically challenge WACSAW with these situations to improve the interpretation of activity data. We assessed 22 unique individuals and repeated the experiment on a large subset of those individuals, and WACSAW performed well across all of those individuals. One participant (11) was sick over one night of the test period. It reduced the high accuracy seen across the other participants and highlights the unique circumstances that exist in individuals with illness or sleep disorders. We anticipate that alterations described in the paragraph above could optimize WACSAW performance across these situations. Also, the vast majority of our population was affiliated with a university, which may inadvertently skew WACSAW results in an unidentified way. Lastly at this time, WACSAW does not output some of the typical metrics provided by other algorithms, which is a future goal of this project so that the program is as useful as possible for the sleep community. It is possible that WACSAW could be used in combination with such analysis suites as GGIR to further improve the understanding of how naturalistic sleep impacts health and behavior.

Despite the above drawbacks to this study, we propose that WACSAW may contribute to classifying sleep in naturalistic settings. WACSAW may be more automated in its application than other methodologies. We submit that WACSAW removes the absolute necessity for participant logs in certain types of studies. In fact, we found circumstances where the segmentation points from WACSAW were more reliable than the participant logs because the log times misidentify behaviors because of memory or attention failures from the participants. This may account for a 1-2% discrepancy based on our data. This makes WACSAW more useful in broad population experiments because there is less human intervention required. Additionally, because WACSAW uses only a single stream of data, namely movement data, to classify behavioral states, longer recordings may be possible based on equal battery life and data storage capacities. Moreover, WACSAW classifies varying length segments rather than fixed length epochs, thereby providing more information that could be utilized to improve the interpretation of a sleep or wakefulness call. The interim metrics may offer opportunities to further identify important patterns and relationships within the movement data. They could be used to determine reliability of the classification results based on these interim metrics. WACSAW adapts to individuals by utilizing the Wasserstein threshold and the characteristic function, both derived from the movement data of that specific individual. In contrast to machine learning algorithms, WACSAW is more explainable while maintaining high accuracy. Lastly, WACSAW and its code will be made available to researchers and the public to evaluate and modify to implement in their particular situation. The code for WACSAW can be found at [45].

Supporting information

S1 Appendix. Results of the hyperparameter grid search..

Acknowledgements

The authors would like to thank the participants for volunteering their time and effort to making this study possible. We would like to thank Don Wunsch II at Missouri S&T for resources and valuable initial discussions. We acknowledge Amber Comfort for her work in preprocessing data for log validation, and Ariel Don for his hard work in creating a GUI version of WACSAW, making it accessible without the difficulties of navigating primary programming languages, as well as for his efforts in optimizing the code. We would like to thank Ozark Biomedical Initiative between Phelps Health Hospitals and Missouri S&T for seed funding to support this project.

References

1. Kochanek KD, Murphy SL, Xu J, Arias E. Mortality in the United States, 2013. NCHS data brief. 2014;(178):1–8.

2. Gottlieb DJ, Punjabi NM, Newman AB, Resnick HE, Redline S, Baldwin CM, et al. Association of Sleep Time With Diabetes Mellitus and Impaired Glucose Tolerance. Archives of Internal Medicine. 2005;165(8):863. doi:10.1001/archinte.165.8.863.

3. Spiegel K, Leproult R, Van Cauter E. Impact of sleep debt on metabolic and endocrine function. The Lancet. 1999;354(9188):1435–1439. doi:10.1016/S0140-6736(99)01376-8.

4. Cappuccio FP, Stranges S, Kandala NB, Miller MA, Taggart FM, Kumari M, et al. Gender-Specific Associations of Short Sleep Duration With Prevalent and Incident Hypertension. Hypertension. 2007;50(4):693–700. doi:10.1161/HYPERTENSIONAHA.107.095471.

5. Ikehara S, Iso H, Date C, Kikuchi S, Watanabe Y, Wada Y, et al. Association of Sleep Duration with Mortality from Cardiovascular Disease and Other Causes for Japanese Men and Women: the JACC Study. Sleep. 2009;32(3):295–301. doi:10.1093/sleep/32.3.295.

6. Mullington JM, Haack M, Toth M, Serrador JM, Meier-Ewert HK. Cardiovascular, Inflammatory, and Metabolic Consequences of Sleep Deprivation. Progress in Cardiovascular Diseases. 2009;51(4):294–302. doi:10.1016/j.pcad.2008.10.003.

7. Dinges DF, Pack F, Williams K, Gillen KA, Powell JW, Ott GE, et al. Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4-5 hours per night. Sleep. 1997;20(4):267–77.

8. Besedovsky L, Lange T, Haack M. The Sleep-Immune Crosstalk in Health and Disease. Physiological Reviews. 2019;99(3):1325–1380. doi:10.1152/physrev.00010.2018.

9. Frey DJ, Fleshner M, Wright KP. The effects of 40 hours of total sleep deprivation on inflammatory markers in healthy young adults. Brain, Behavior, and Immunity. 2007;21(8):1050–1057. doi:10.1016/j.bbi.2007.04.003.

10. Mullington JM, Simpson NS, Meier-Ewert HK, Haack M. Sleep loss and inflammation. Best Practice & Research Clinical Endocrinology & Metabolism. 2010;24(5):775–784. doi:10.1016/j.beem.2010.08.014.

11. Cappuccio FP, D'Elia L, Strazzullo P, Miller MA. Sleep Duration and All-Cause Mortality: A Systematic Review and Meta-Analysis of Prospective Studies. Sleep. 2010;33(5):585–592. doi:10.1093/sleep/33.5.585.

12. Gallicchio L, Kalesan B. Sleep duration and mortality: a systematic review and meta-analysis. Journal of Sleep Research. 2009;18(2):148–158. doi:10.1111/j.1365-2869.2008.00732.x.

13. Grandner MA, Hale L, Moore M, Patel NP. Mortality associated with short sleep duration: The evidence, the possible mechanisms, and the future. Sleep Medicine Reviews. 2010;14(3):191–203. doi:10.1016/j.smrv.2009.07.006.

14. Heslop P, Smith GD, Metcalfe C, Macleod J, Hart C. Sleep duration and mortality: the effect of short or long sleep duration on cardiovascular and all-cause mortality in working men and women. Sleep Medicine. 2002;3(4):305–314. doi:10.1016/S1389-9457(02)00016-3.

15. Kripke DF, Garfinkel L, Wingard DL, Klauber MR, Marler MR. Mortality Associated With Sleep Duration and Insomnia. Archives of General Psychiatry. 2002;59(2):131. doi:10.1001/archpsyc.59.2.131.

16. Carney CE, Lajos LE, Waters WF. Wrist Actigraph Versus Self-Report in Normal Sleepers: Sleep Schedule Adherence and Self-Report Validity. Behavioral Sleep Medicine. 2004;2(3):134–143. doi:10.1207/s15402010bsm0203_2.

17. Van Den Berg JF, Van Rooij FJA, Vos H, Tulen JHM, Hofman A, Miedema HME, et al. Disagreement between subjective and actigraphic measures of sleep duration in a population-based study of elderly persons*. Journal of Sleep Research. 2008;17(3):295–302. doi:10.1111/j.1365-2869.2008.00638.x.

18. Basner M, Dinges DF. Dubious Bargain: Trading Sleep for Leno and Letterman. Sleep. 2009;32(6):747–752. doi:10.1093/sleep/32.6.747.

19. Chattu VK, Manzar MD, Kumary S, Burman D, Spence DW, Pandi-Perumal SR. The Global Problem of Insufficient Sleep and Its Serious Public Health Implications. Healthcare (Basel, Switzerland). 2018;7(1). doi:10.3390/healthcare7010001.

20. Ding L, Chen B, Dai Y, Li Y. A meta-analysis of the first-night effect in healthy individuals for the full age spectrum. Sleep Medicine. 2022;89:159–165. doi:10.1016/j.sleep.2021.12.007.

21. Morgenthaler T, Alessi C, Friedman L, Owens J, Kapur V, Boehlecke B, et al. Practice Parameters for the Use of Actigraphy in the Assessment of Sleep and Sleep Disorders: An Update for 2007. Sleep. 2007;30(4):519–529. doi:10.1093/sleep/30.4.519.

22. Sadeh A, Alster J, Urbach D, Lavie P. Actigraphically based automatic bedtime sleep-wake scoring: validity and clinical applications. Journal of Ambulatory Monitoring. 1989;.

23. 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.

24. Van de Water ATM, Holmes A, Hurley DA. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography – a systematic review. Journal of Sleep Research. 2011;20(1pt2):183–200. doi:10.1111/j.1365-2869.2009.00814.x.

25. Palotti J, Mall R, Aupetit M, Rueschman M, Singh M, Sathyanarayana A, et al. Benchmark on a large cohort for sleep-wake classification with machine learning techniques. npj Digital Medicine. 2019;2(1):50. doi:10.1038/s41746-019-0126-9.

26. Khademi A, EL-Manzalawy Y, Master L, Buxton OM, Honavar VG. Personalized Sleep Parameters Estimation from Actigraphy: A Machine Learning Approach. Nature and Science of Sleep. 2019;Volume 11:387–399. doi:10.2147/NSS.S220716.

27. Cho T, Sunarya U, Yeo M, Hwang B, Koo YS, Park C. Deep-ACTINet: End-to-End Deep Learning Architecture for Automatic Sleep-Wake Detection Using Wrist Actigraphy. Electronics. 2019;8(12):1461. doi:10.3390/electronics8121461.

28. Cakmak AS, Da Poian G, Willats A, Haffar A, Abdulbaki R, Ko YA, et al. An unbiased, efficient sleep–wake detection algorithm for a population with sleep disorders: change point decoder. Sleep. 2020;43(8). doi:10.1093/sleep/zsaa011.

29. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Performance comparison of different interpretative algorithms utilized to derive sleep parameters from wrist actigraphy data. Chronobiology International. 2019;36(12):1752–1760. doi:10.1080/07420528.2019.1679826.

30. Jansen EC, Dunietz GL, Tsimpanouli ME, Guyer HM, Shannon C, Hershner SD, et al. Sleep, Diet, and Cardiometabolic Health Investigations: a Systematic Review of Analytic Strategies. Current Nutrition Reports. 2018;7(4):235–258. doi:10.1007/s13668-018-0240-3.

31. Buysse DJ, Cheng Y, Germain A, Moul DE, Franzen PL, Fletcher M, et al. Night-to-night sleep variability in older adults with and without chronic insomnia. Sleep Medicine. 2010;11(1):56–64. doi:10.1016/j.sleep.2009.02.010.

32. Mezick EJ, Matthews KA, Hall M, Kamarck TW, Buysse DJ, Owens JF, et al. Intra-individual variability in sleep duration and fragmentation: Associations with stress. Psychoneuroendocrinology. 2009;34(9):1346–1354. doi:10.1016/j.psyneuen.2009.04.005.

33. Migueles JH, Rowlands AV, Huber F, Sabia S, van Hees VT. GGIR: A Research Community–Driven Open Source R Package for Generating Physical Activity and Sleep Outcomes From Multi-Day Raw Accelerometer Data. Journal for the Measurement of Physical Behaviour. 2019;2(3):188–196. doi:10.1123/jmpb.2018-0063.

34. Van Hees VT, Sabia S, Jones SE, Wood AR, Anderson KN, Kivimäki M, et al. Estimating sleep parameters using an accelerometer without sleep diary. Scientific Reports. 2018;8(1):12975. doi:10.1038/s41598-018-31266-z.

35. Actiware. Actiware Tutorials; 2019. Available from: https://www.usa.philips.com/healthcare/sites/actigraphy/solutions/actiware.

36. GENEActiv. GENEActiv Instructions v1.4. Activeinsights Ltd.; 2019.

37. Murphy SL. Review of physical activity measurement using accelerometers in older adults: Considerations for research design and conduct. Preventive Medicine. 2009;48(2):108–114. doi:10.1016/j.ypmed.2008.12.001.

38. Van Hees VT, Sabia S, Anderson KN, Denton SJ, Oliver J, Catt M, et al. A Novel, Open Access Method to Assess Sleep Duration Using a Wrist-Worn Accelerometer. PLOS ONE. 2015;10(11):e0142533. doi:10.1371/journal.pone.0142533.

39. Villani C. Topics in Optimal Transportation. American Mathematical Society; 2003.

40. Villani C. Optimal Transport: Old and New. vol. 338 of Grundlehren der mathematischen Wissenschaften. Berlin, Heidelberg: Springer Berlin Heidelberg; 2009.

41. Peyré G, Cuturi M. Computational Optimal Transport. 1st ed. Now Publishers Inc.; 2019.

42. Levene H. Robust Tests for Equality of Variances. In: Olkin I, editor. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. 1st ed. Stanford University Press; 1960. p. 278–292.

43. Bitkina OV, Park J, Kim J. Modeling Sleep Quality Depending on Objective Actigraphic Indicators Based on Machine Learning Methods. International Journal of Environmental Research and Public Health. 2022;19(16):9890. doi:10.3390/ijerph19169890.

44. Li X, Zhang Y, Jiang F, Zhao H. A novel machine learning unsupervised algorithm for sleep/wake identification using actigraphy. Chronobiology International. 2020;37(7):1002–1015. doi:10.1080/07420528.2020.1754848.

45. Vandegriffe A, Don A. WACSAW; 2023. **Pending release**.

Hyperparameter Grid Search

The final WACSAW algorithm contains 4 hyperparameters to which we assigned values as we developed the WACSAW process. These assignments were based on quick trial and error, but they needed to be formally tuned to determine what values optimize WACSAW output. We used data from the same 6 training individuals with their varying sleep characteristics and tuned the model using accuracy to determine the hyperparameters that led to the best outcomes.

We needed to tune (1) the order of the Wasserstein distance (p in Eq. 3); (2) the window size of data input; (3) the change point threshold for segmentation; and (4) the significance value of the Levene test.

For the order of the Wasserstein distance, we considered $p \in \{1, 2\}$. For the window size, we considered windows of size 2, 5, 7 minutes. For the change point threshold, we considered the statistics mean, median, and α-trimmed mean with $\alpha = 0.25$ for a 2-day period. For the Levene significant level, we checked significance levels between 0 and 1 ($\{10^{-s} : s \in [0, 81]\}$) (note that the extremely small values were computed using arbitrary precision arithmetic). It may be more appropriate to treat this value, not as a significance level of a test, but as a parameter one could tune to accentuate the differences between sleep and wakeful segments, such as quiescent wakefulness.

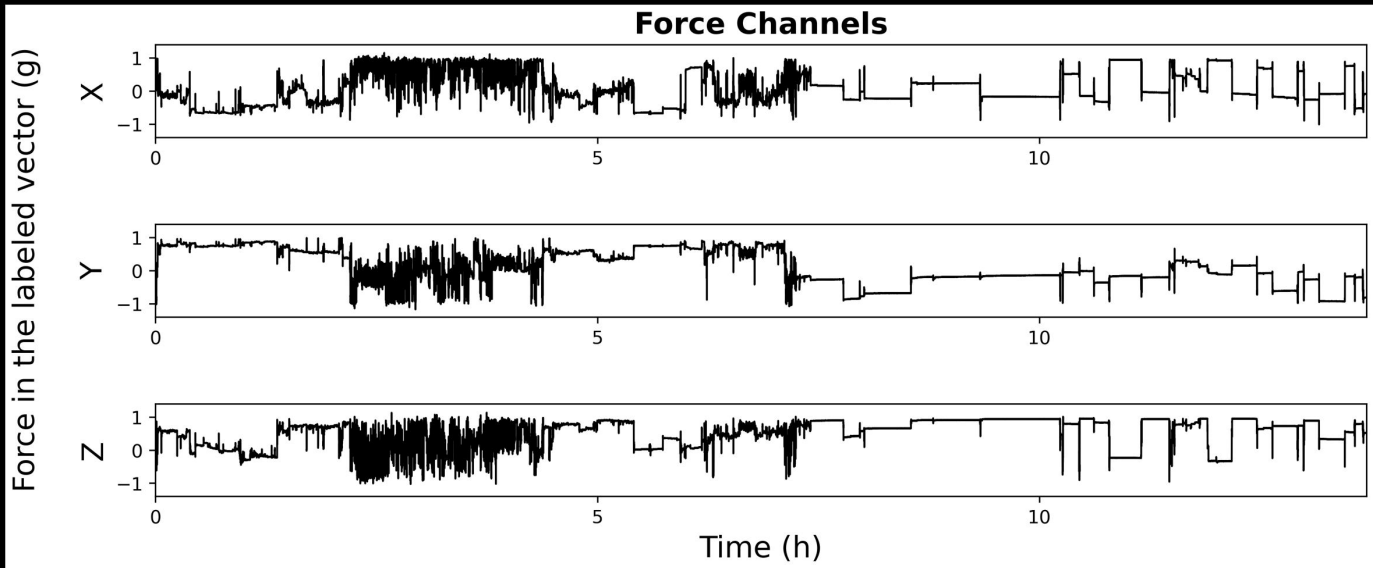
We use parameter sweeps to test the hyperparameters. This revealed that WACSAW is not overly sensitive to variations in the hyperparameters within the search space,

indicating that inadvertent deviations from optimal do not change the final output of WACSAW to any fundamental extent. Comparing the accuracy of WACSAW to the adjusted log classifications revealed that the accuracy stayed above 90% for all hyperparameter combinations on the training set and standard deviation ranged from 1 to below 11, computed from the 6 individuals in the development set. Moreover, the standard deviation decreases as accuracy increases. The top performing combination was order 1 with a 5 minute window, a mean change point threshold, and a Levene significance level of 1e-25 (Table 6). But the top ten combinations all had an accuracy above 97% but all combinations also had a standard deviation in accuracy of 1.30-2.8% between individuals, which indicates that the adaptive aspects of WACSAW reduced the variability in accuracy sometimes observed when a common algorithm is applied between individuals.

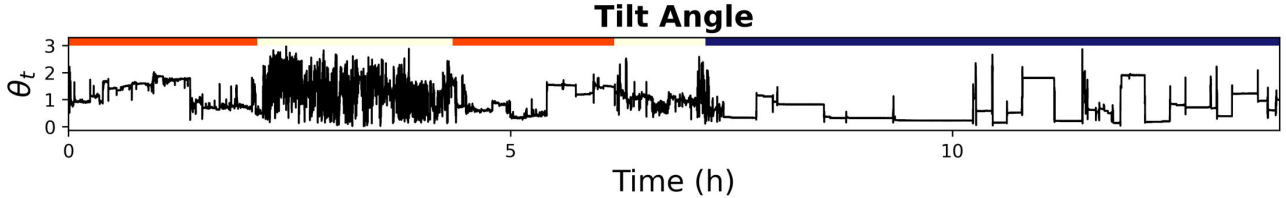
WTS Order	WTS Window Size	CPT Function	Levene Threshold	% Correct Accuracy (Raw)	% Correct Sensitivity (Raw)	% Correct Specificity (Raw)	% Correct Accuracy (Adjusted)	% Correct Sensitivity (Adjusted)	% Correct Specificity (Adjusted)
1	300	mean	1.00E-25	96.8	96.3	97.1	97.7	97.6	97.7
-	-	-	-	1.2	2.1	1.6	1.5	1.8	1.5
1	480	mean	1.00E-40	96.7	96.7	96.8	97.6	98.0	97.5
-	-	-	-	1.1	1.8	1.6	1.3	2.1	1.4
1	480	mean	1.00E-45	96.7	96.7	96.8	97.6	98.0	97.5
-	-	-	-	1.1	1.8	1.6	1.3	2.1	1.4
1	480	mean	1.00E-30	96.7	96.7	96.8	97.6	98.0	97.4
-	-	-	-	1.3	1.6	1.8	1.3	2.1	1.3
1	300	mean	1.00E-35	96.7	95.1	97.3	97.4	96.1	97.8
-	-	-	-	1.6	6.4	1.6	1.9	5.6	1.3
1	480	mean	1.00E-25	96.6	96.5	96.8	97.5	97.7	97.4
-	-	-	-	1.3	1.9	1.9	1.3	2.0	1.3
1	480	mean	1.00E-20	96.6	96.3	96.9	97.4	97.5	97.5
-	-	-	-	1.3	1.8	1.8	1.4	1.9	1.3
1	480	mean	1.00E-15	96.5	96.2	96.8	97.4	97.4	97.4
-	-	-	-	1.2	2.0	1.7	1.5	1.9	1.4
1	300	mean	1.00E-30	96.5	96.4	96.6	97.3	97.8	97.2
-	-	-	-	1.9	1.9	2.7	1.7	1.8	1.7
1	480	mean	1.00E-50	96.4	96.7	96.4	97.4	98.1	97.1
-	-	-	-	2.0	1.8	2.9	1.5	2.1	1.8
1	480	mean	1.00E-55	96.4	96.7	96.4	97.4	98.1	97.1
-	-	-	-	2.0	1.8	2.9	1.5	2.1	1.8
1	300	mean	1.00E-20	96.4	95.6	96.9	97.3	96.9	97.5
-	-	-	-	1.6	3.7	1.5	1.6	2.2	1.5
1	480	mean	1.00E-60	96.4	96.5	96.4	97.3	98.0	97.1
-	-	-	-	2.0	2.3	2.9	1.5	2.2	1.8
1	300	median	1.00E-35	96.3	95.0	97.1	97.2	96.0	97.7
-	-	-	-	2.1	5.8	2.0	1.8	3.0	1.3
1	480	mean	1.00E-05	96.3	95.6	96.8	97.2	96.8	97.4
-	-	-	-	1.3	3.6	1.8	1.8	2.4	1.4
1	480	mean	1.00E+00	96.3	95.9	96.5	97.1	97.2	97.1
-	-	-	-	1.2	2.6	1.9	1.2	1.7	1.3
2	300	mean	1.00E-60	96.3	93.9	97.6	97.1	94.7	98.2
-	-	-	-	3.1	10.0	1.4	2.8	5.5	1.4
1	300	median	1.00E-40	96.3	94.5	97.3	97.1	95.4	97.9
-	-	-	-	2.7	8.5	2.1	2.3	4.5	1.4
1	300	median	1.00E-30	96.2	94.7	97.0	97.0	95.7	97.6
-	-	-	-	2.2	6.6	2.0	1.9	3.4	1.3
1	300	median	1.00E-25	96.1	94.5	97.1	97.0	95.5	97.7
-	-	-	-	2.1	7.0	1.6	1.9	3.6	1.2

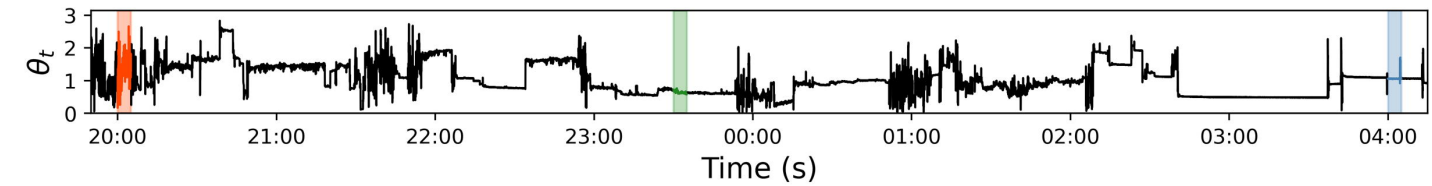
Table 6. Top 20 results based on accuracy from the hyperparameter sweep. The initial columns are the values of the hyperparameters and the validation metric columns are the metrics for WACSAW when run with the hyperparameter combination on the development cohort.

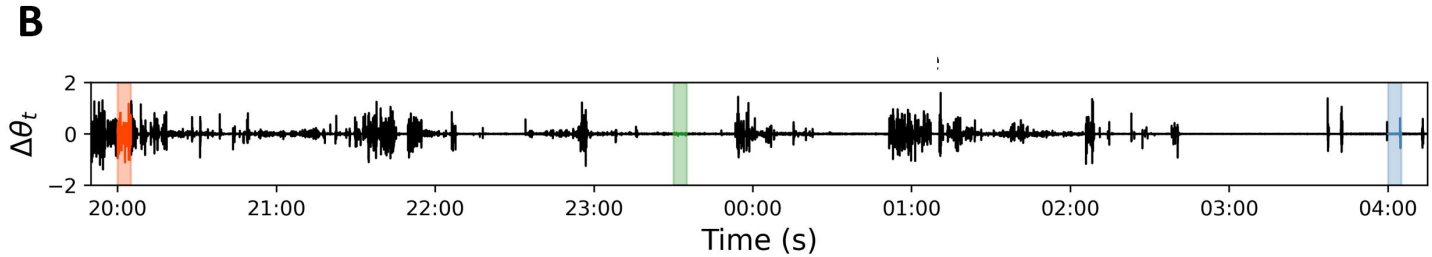
Other hyperparameters include the permutation test threshold, the ξ cutoff for the characteristic functions, and the characteristic clustering parameter, namely the 90% closeness parameter. These parameters were not subject to tuning as they were either empirically chosen, restricted due to computation, or were heuristically chosen.

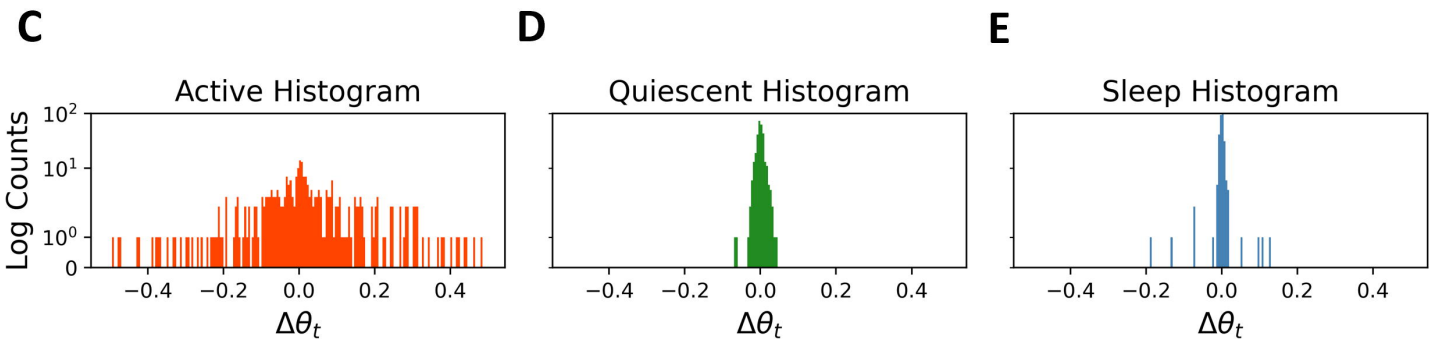


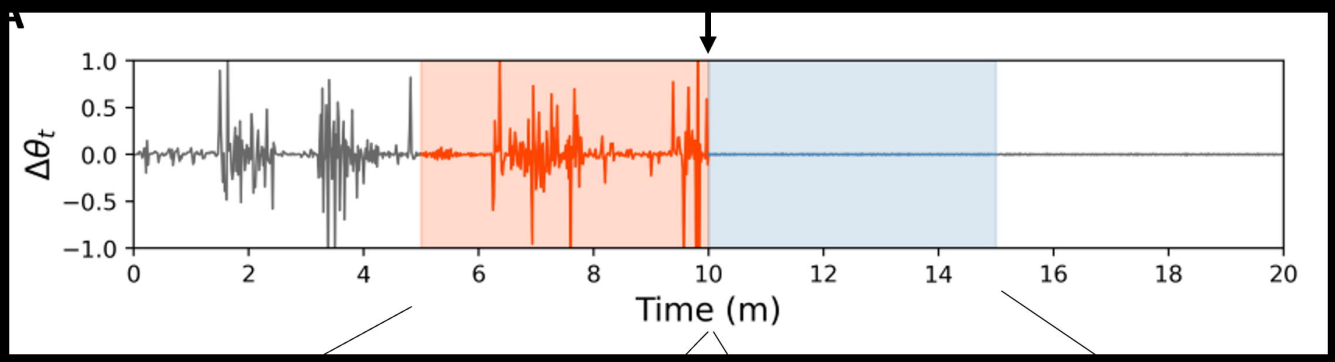
$$\theta_t = \arccos\left(\frac{z_t}{\sqrt{x_t^2 + y_t^2 + z_t^2}}\right)$$

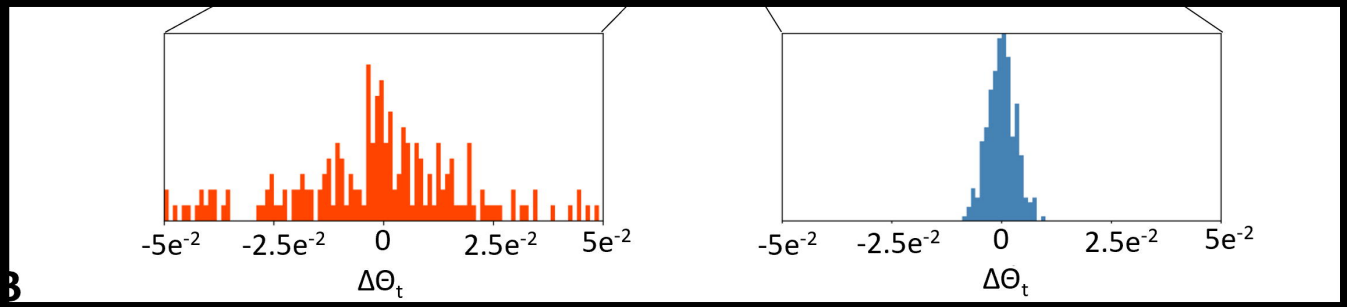


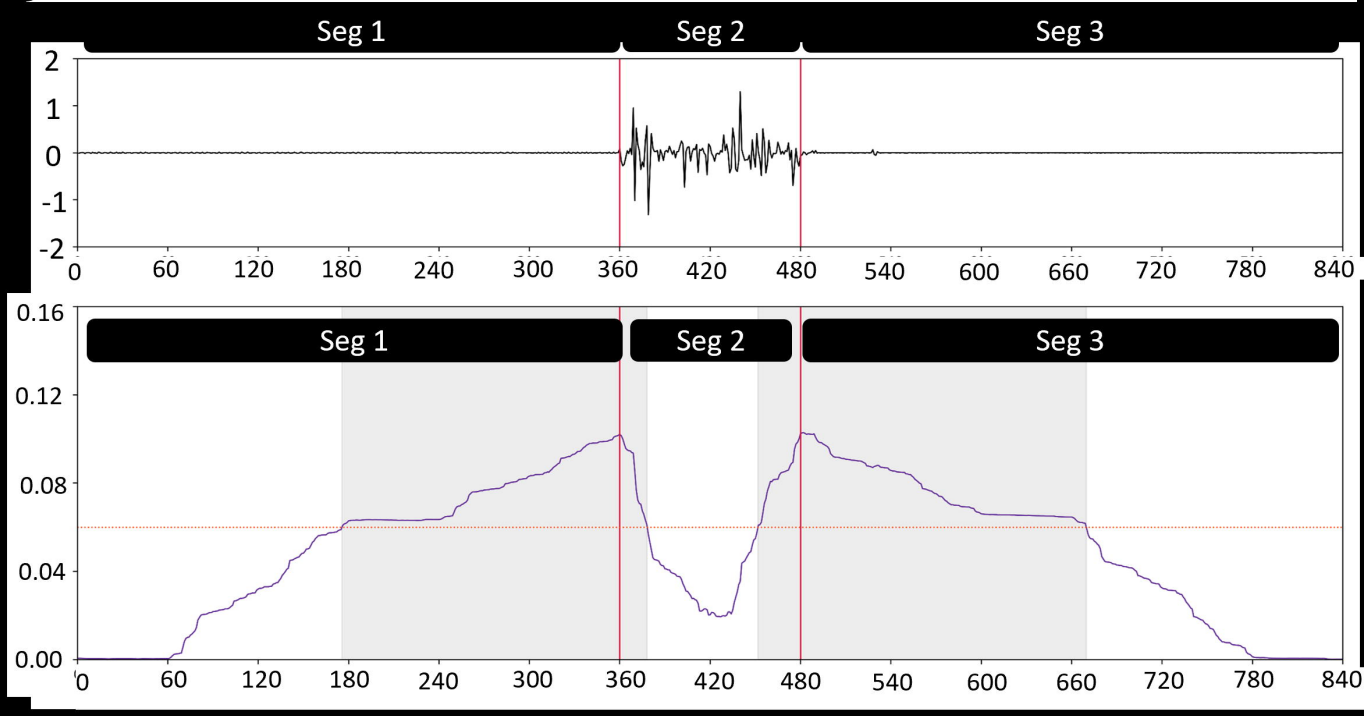

θ_t
3
2
1
0
20:00
21:00
22:00
23:00
00:00
01:00
02:00
03:00
04:00
Time (s)

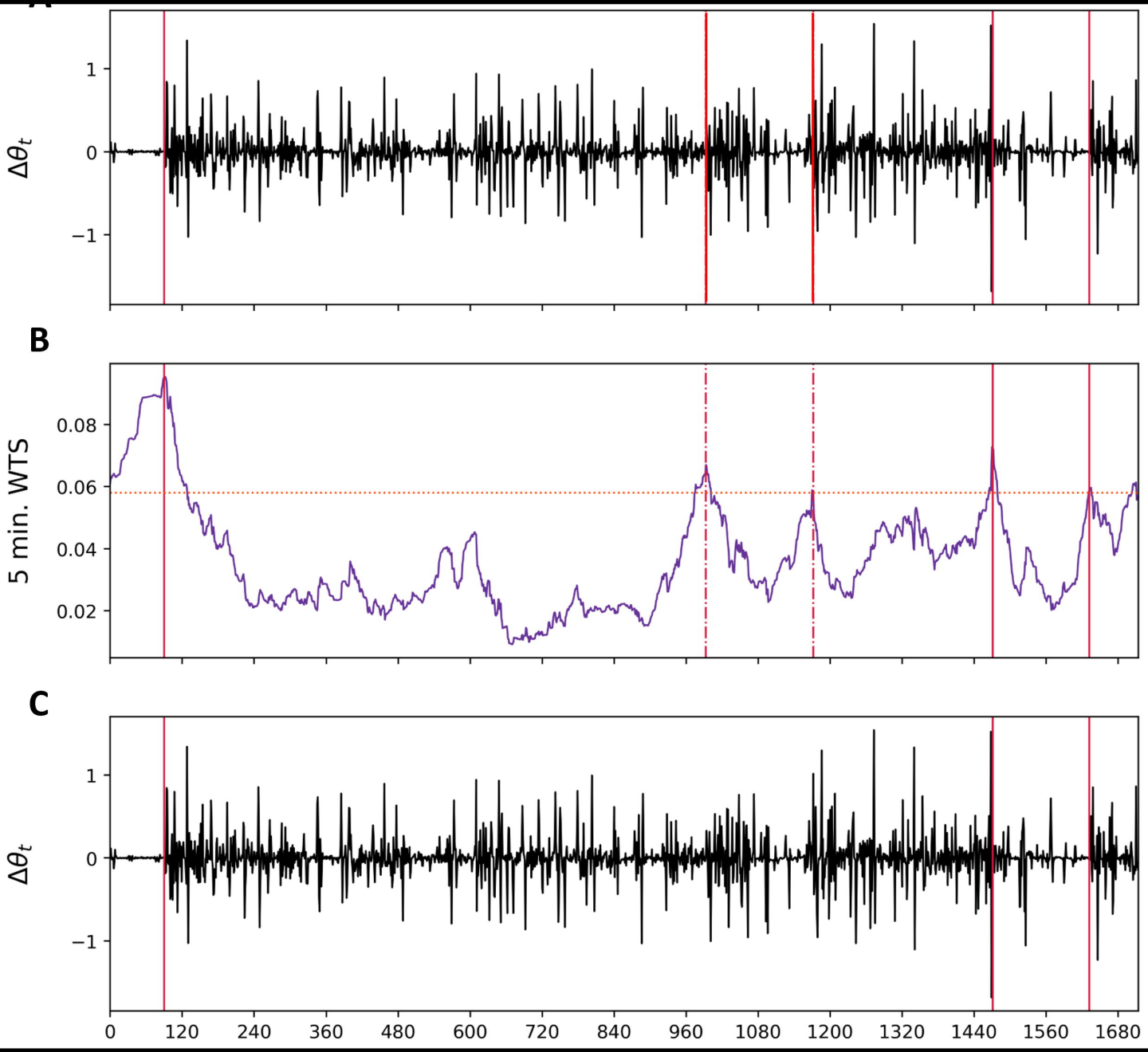

B
$\Delta\theta_t$
2
0
−2
20:00
21:00
22:00
23:00
00:00
01:00
02:00
03:00
04:00
Time (s)

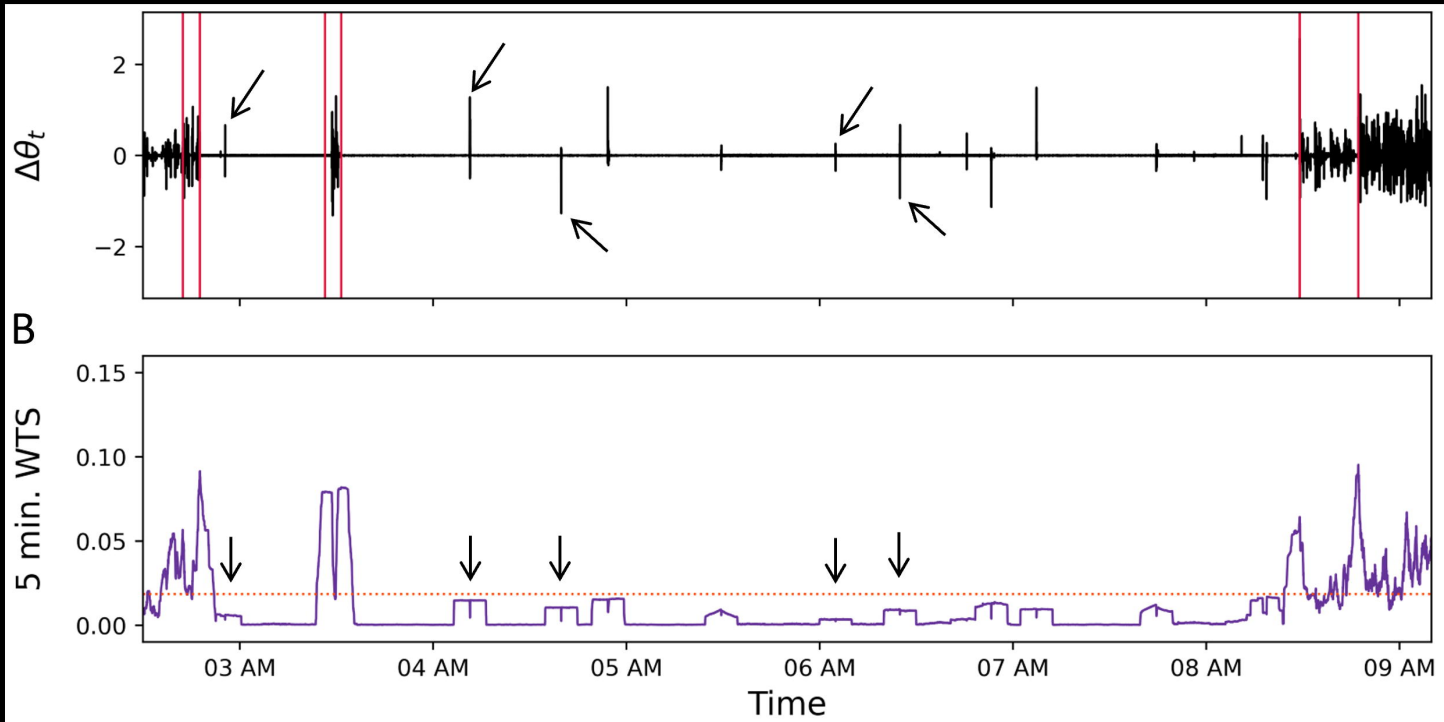

C
D
E
Active Histogram
Quiescent Histogram
Sleep Histogram
Log Counts
10^2
10^1
10^0
0
−0.4
−0.2
0.0
0.2
0.4
$\Delta\theta_t$

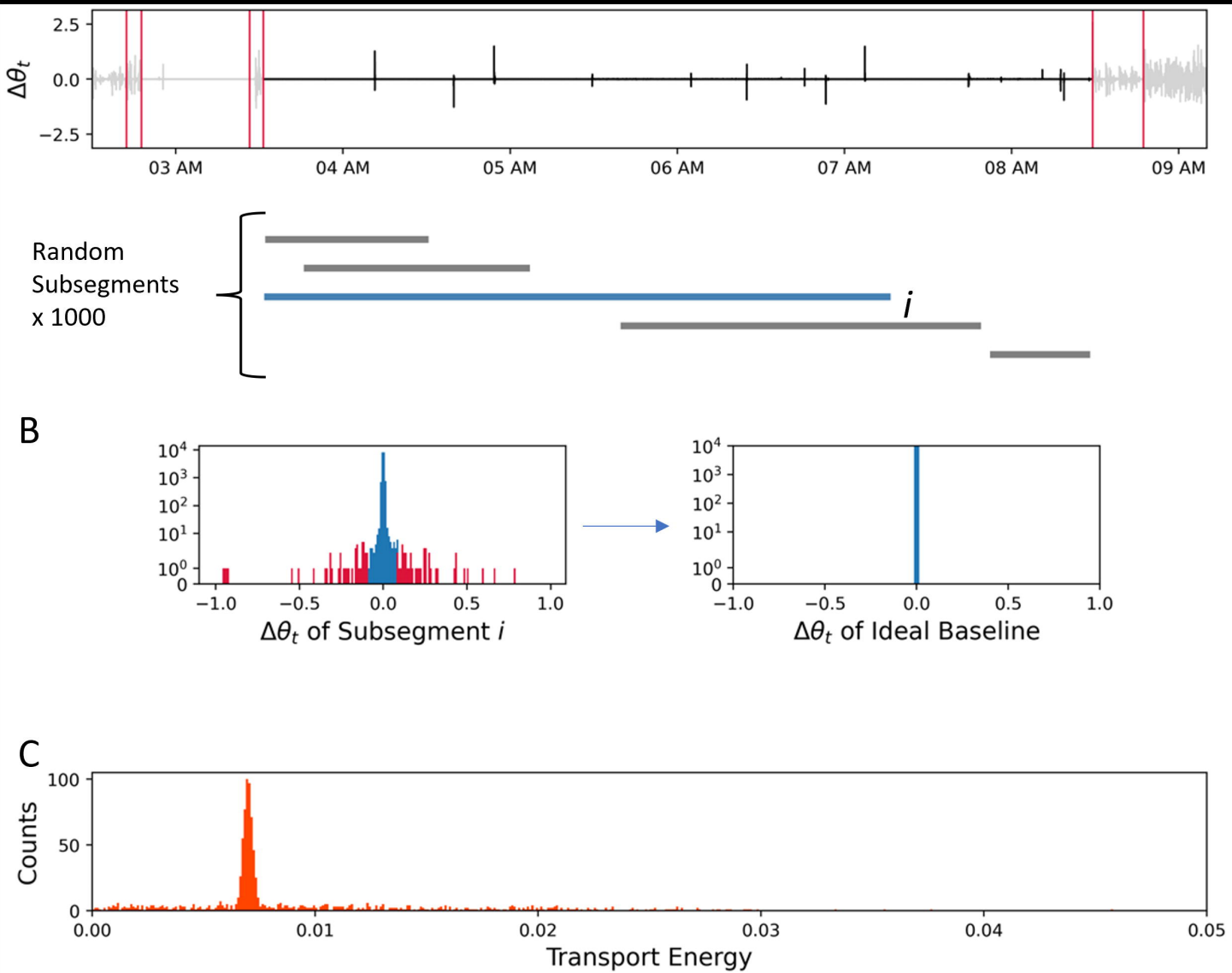

A
1.0
0.5
0.0
−0.5
−1.0
$\Delta\theta_t$
0
2
4
6
8
10
12
14
16
18
20
Time (m)

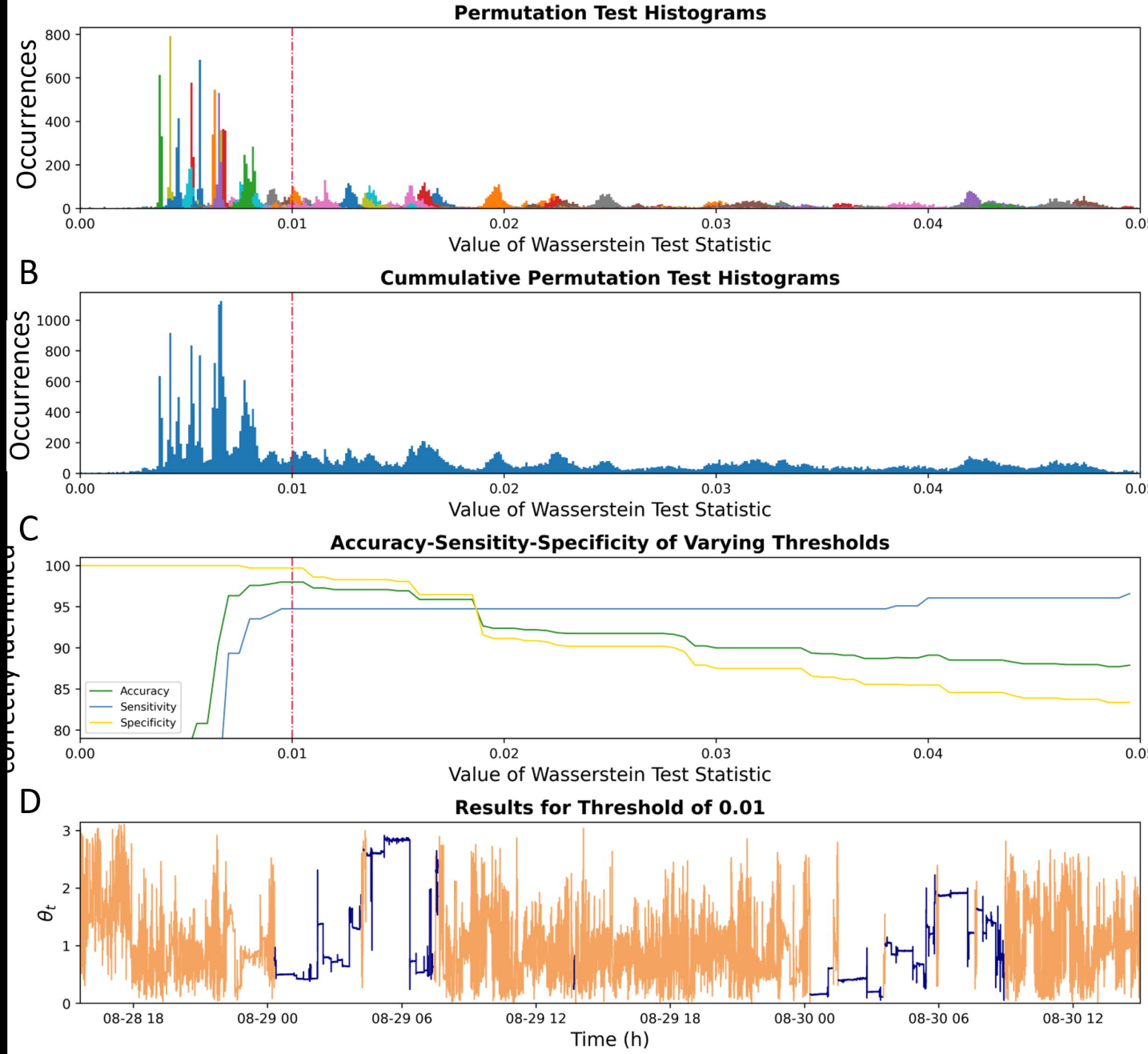

-5e-2
-2.5e-2
0
2.5e-2
5e-2
$\Delta\Theta_t$
-5e-2
-2.5e-2
0
2.5e-2
5e-2
$\Delta\Theta_t$

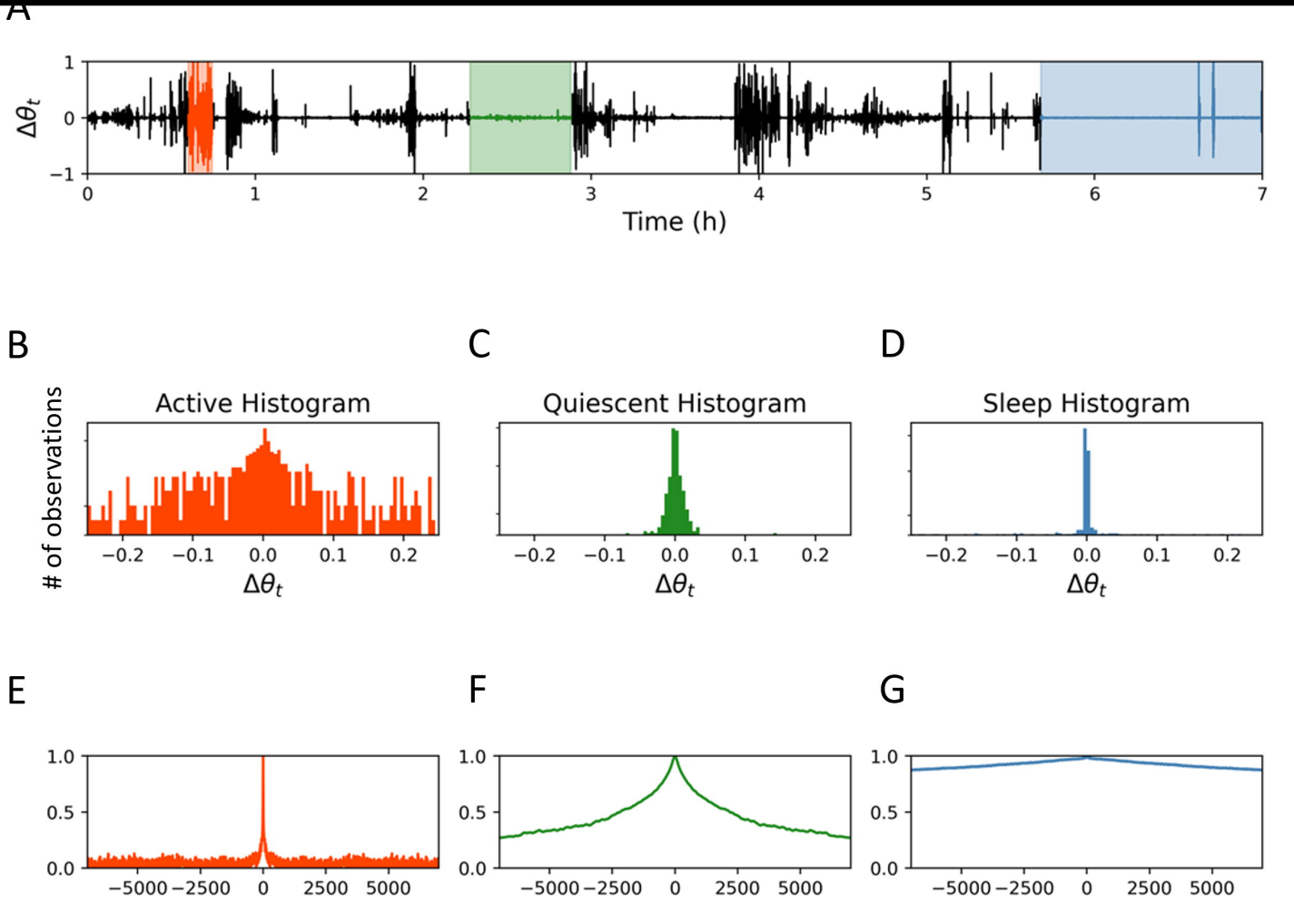

B
Seg 1
Seg 2
Seg 3
2
1
0
-1
-2
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840
Seg 1
Seg 2
Seg 3
0.16
0.12
0.08
0.04
0.00
0
60
120
180
240
300
360
420
480
540
600
660
720
780
840

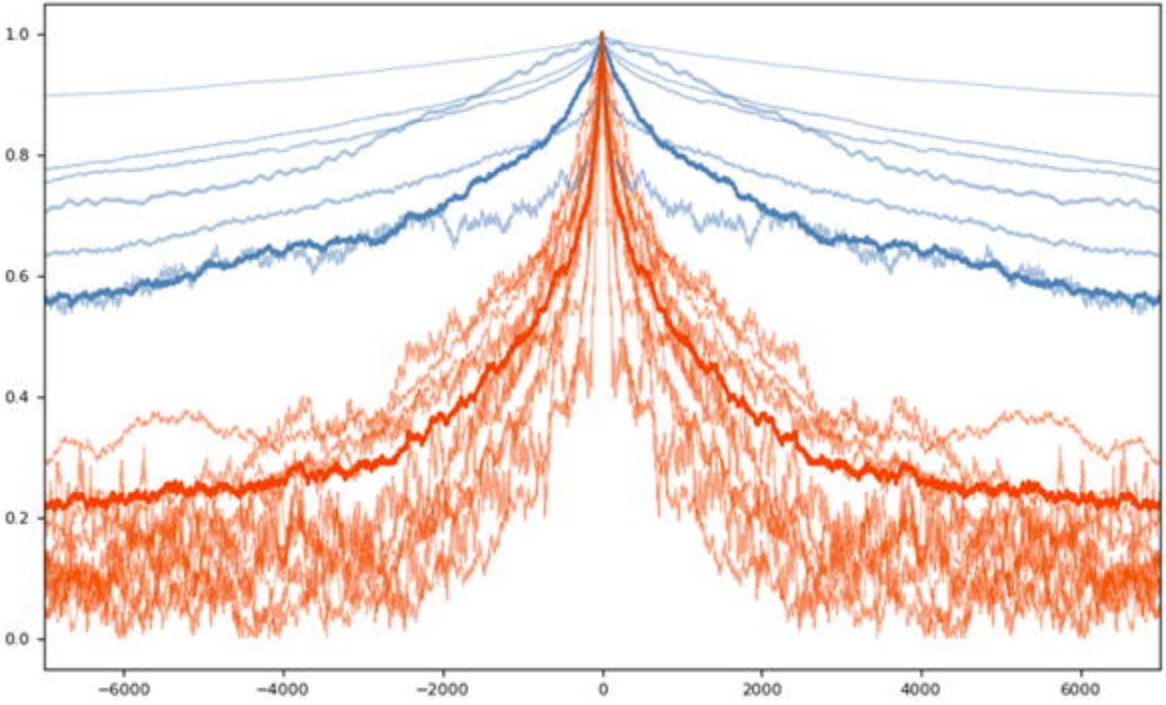

A
B
C
$\Delta\theta_t$
5 min. WTS
1
0
−1
0.08
0.06
0.04
0.02
0
120
240
360
480
600
720
840
960
1080
1200
1320
1440
1560
1680


2
0
−2
$\Delta\theta_t$
B
0.15
0.10
0.05
0.00
5 min. WTS
03 AM
04 AM
05 AM
06 AM
07 AM
08 AM
09 AM
Time

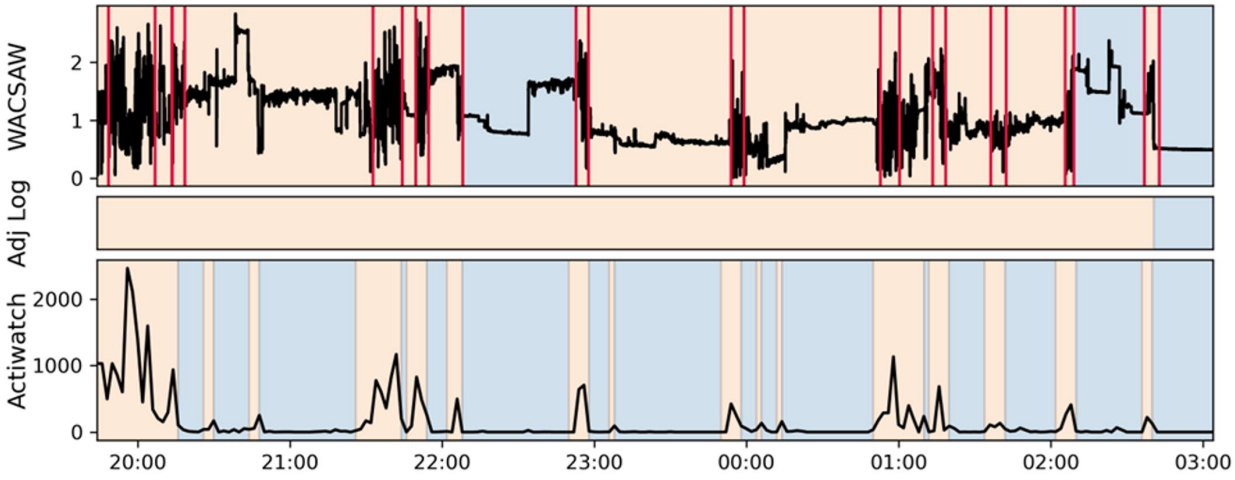

$\Delta\theta_t$
2.5
0.0
−2.5
03 AM
04 AM
05 AM
06 AM
07 AM
08 AM
09 AM
Random
Subsegments
x 1000
i
B
10^4
10^3
10^2
10^1
10^0
0
−1.0
−0.5
0.0
0.5
1.0
$\Delta\theta_t$ of Subsegment i
$\Delta\theta_t$ of Ideal Baseline
C
Counts
100
50
0
0.00
0.01
0.02
0.03
0.04
0.05
Transport Energy


Permutation Test Histograms
Occurrences
Value of Wasserstein Test Statistic
B
Cummulative Permutation Test Histograms
Occurrences
Value of Wasserstein Test Statistic
C
Accuracy-Sensitity-Specificity of Varying Thresholds
Accuracy
Sensitivity
Specificity
Value of Wasserstein Test Statistic
D
Results for Threshold of 0.01
θ_t
08-28 18
08-29 00
08-29 06
08-29 12
08-29 18
08-30 00
08-30 06
08-30 12
Time (h)

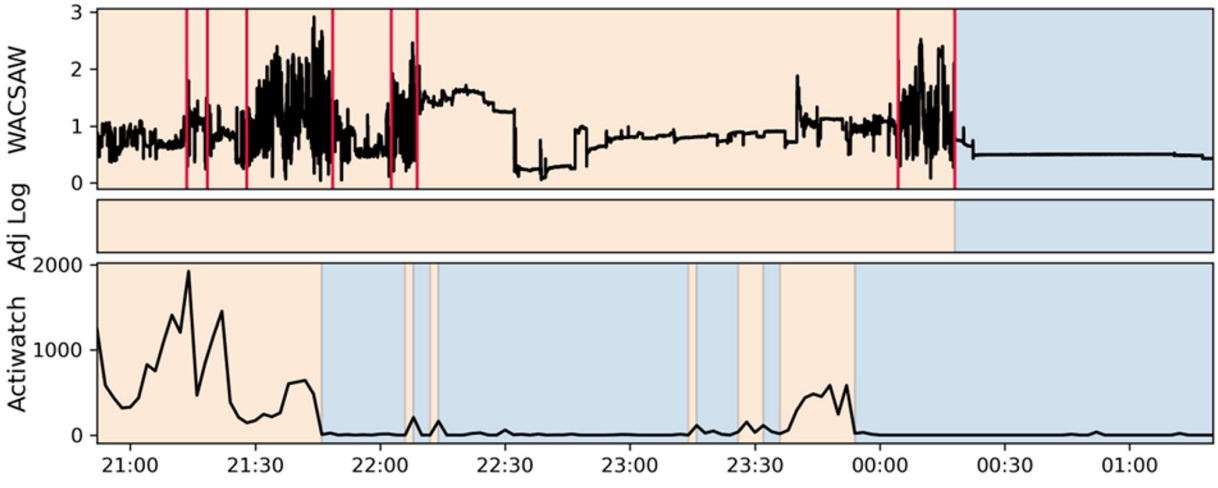

A
$\Delta\theta_t$
Time (h)
B
Active Histogram
# of observations
$\Delta\theta_t$
C
Quiescent Histogram
$\Delta\theta_t$
D
Sleep Histogram
$\Delta\theta_t$
E
F
G


1.0
0.8
0.6
0.4
0.2
0.0
−6000
−4000
−2000
0
2000
4000
6000

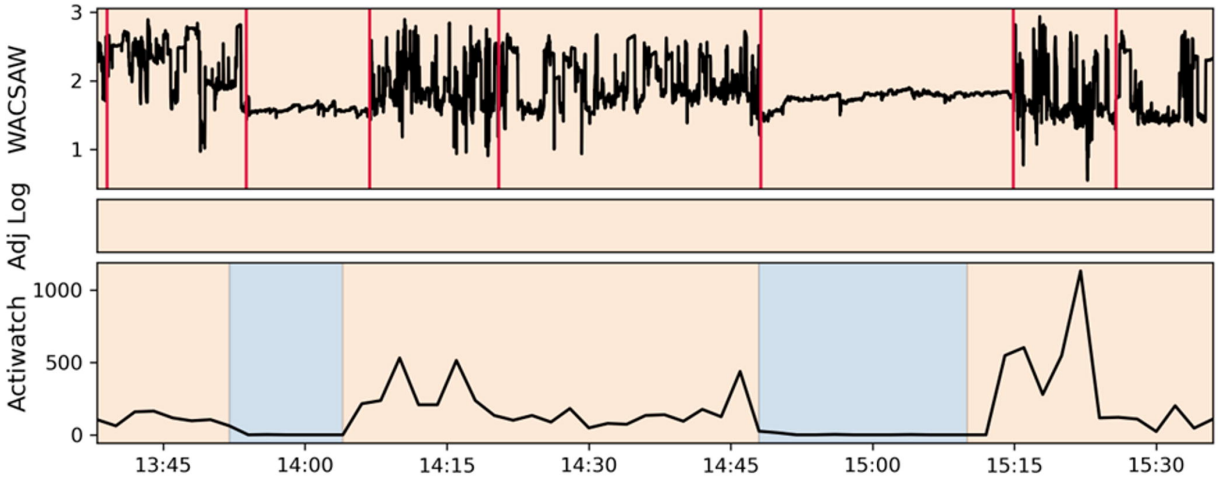
A


WACSAW
Adj Log
Actiwatch
0
1
2
0
1000
2000
20:00
21:00
22:00
23:00
00:00
01:00
02:00
03:00

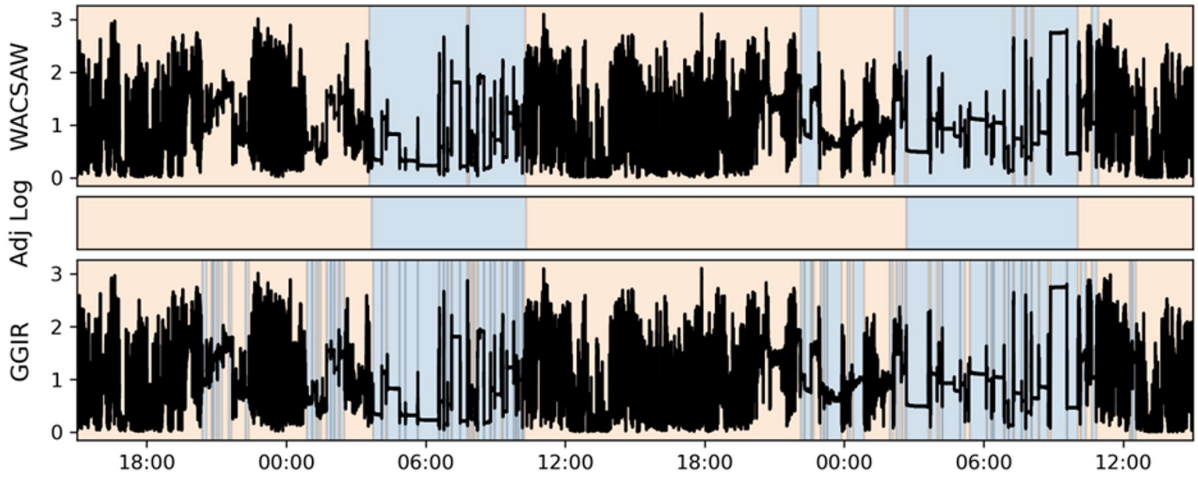
B


WACSAW
Adj Log
Actiwatch
0
1
2
3
0
1000
2000
21:00
21:30
22:00
22:30
23:00
23:30
00:00
00:30
01:00

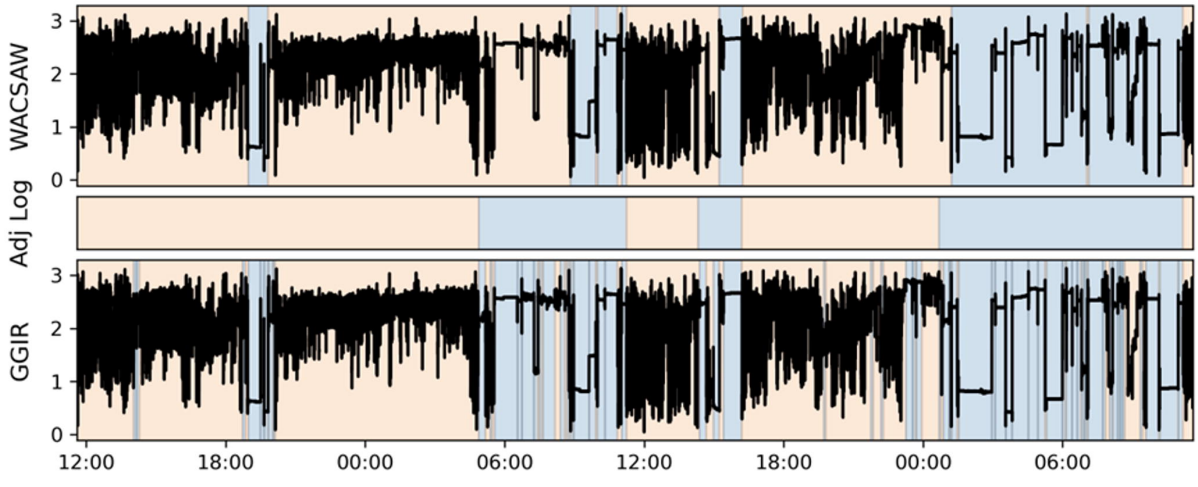
C


WACSAW
Adj Log
Actiwatch
1
2
3
0
500
1000
13:45
14:00
14:15
14:30
14:45
15:00
15:15
15:30

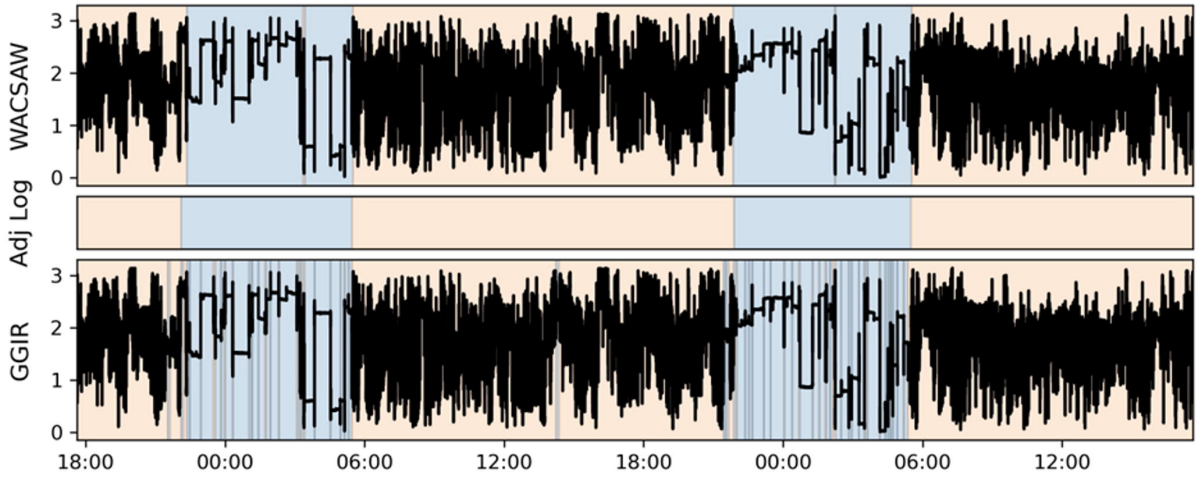
A


WACSAW
Adj Log
GGIR
0
1
2
3
18:00
00:00
06:00
12:00
18:00
00:00
06:00
12:00

B


WACSAW
Adj Log
GGIR
0
1
2
3
12:00
18:00
00:00
06:00
12:00
18:00
00:00
06:00

C


WACSAW
Adj Log
GGIR
0
1
2
3
18:00
00:00
06:00
12:00
18:00
00:00
06:00
12:00

WACAW development	Participant id	% correct Accuracy (Raw)	% correct Sensitivity (raw)	% correct Specificity (raw)	% correct Accuracy (Adjusted)	% correct Sensitivity (Adj)	% correct Selectivity (adj)
Development Cohort	1	96.9	94.6	98.2	95.8	94.3	96.7
	2	96.7	95.9	97.0	98.8	98.7	98.9
	3	99.1	98.4	99.5	99.5	99.5	99.5
	4	96.0	93.3	97.5	97.6	96.8	98.0
	5	95.5	96.5	95.1	95.8	97.2	95.3
	6	96.5	99.0	95.5	98.4	99.0	98.1
	Mean	96.8	96.3	97.1	97.7	97.6	97.7
	Median	96.6	96.2	97.2	98.0	97.9	98.1
	Std Dev	1.2	2.2	1.7	1.5	1.9	1.5
	MAD	0.8	1.7	1.3	1.2	1.5	1.1
	min	95.5	93.3	95.1	95.8	94.3	95.3
	max	99.1	99.0	99.5	99.5	99.5	99.5
	% > 95%	100.0	100.0	100.0	100.0	100.0	100.0
Development Cohort Validation	1	92.4	94.7	91.3	94.4	97.3	93.0
	2	97.2	93.5	98.8	98.1	95.3	99.2
	3	96.6	99.4	95.3	96.4	99.6	95.0
	4	94.4	91.2	96.4	96.8	99.0	95.7
	5	90.6	97.8	87.6	98.7	98.3	98.9
	Mean	94.2	95.3	93.9	96.9	97.9	96.4
	Median	94.4	94.7	95.3	96.8	98.3	95.7
	Std Dev	2.8	3.3	4.4	1.6	1.7	2.7
	MAD	2.1	2.5	3.2	1.2	1.2	
	min	90.6	91.2	87.6	94.4	95.3	93.0
	max	97.2	99.4	98.8	98.7	99.6	99.2
	% > 95%	83.3	83.3	83.3	83.3	83.3	83.3

Std Dev = Standard deviation

MAD = Mean absolute deviation from the median

	Participant id	% correct Accuracy (raw)	% correct Sensitivity (raw)	% correct Specificity (raw)	% correct Accuracy (adj)	% correct Sensitivity (adj)	% correct Selectivity (adj)
Independent Validation	7	97.7	94.2	99.4	99.4	99.2	99.5
	8	96.6	94.4	97.9	97.7	97.3	98.0
	9	97.4	94.4	99.0	98.8	97.1	99.6
	10	97.3	94.8	98.2	97.2	94.7	98.1
	11	85.3	68.4	96.0	86.3	69.4	97.0
	12	98.9	98.3	99.2	98.9	98.4	99.1
	13	96.6	98.6	95.4	98.4	99.5	97.7
	14	95.5	99.6	94.3	96.7	99.9	95.7
	15	95.5	88.1	100.0	98.6	97.2	99.3
	16	86.4	98.3	82.4	98.8	97.2	99.7
	17	98.8	97.9	99.2	98.7	98.0	99.2
	18	97.6	98.8	96.9	98.8	99.3	98.5
	19	99.4	99.1	99.5	99.1	97.7	99.7
	20	98.9	96.8	100.0	98.9	97.0	100.0
	21	94.8	96.7	94.1	97.0	98.5	96.3
	22	86.4	91.8	84.0	89.9	95.5	87.3
	Mean	95.2	94.4	96.0	97.1	96.0	97.8
	Median	97.0	96.7	98.1	98.7	97.5	98.8
	Std Dev	4.7	7.6	5.4	3.7	7.2	3.1
	MAD	3.1	4.0	3.4	1.8	2.8	1.7
	min	85.3	68.4	82.4	86.3	69.4	87.3
	max	99.4	99.6	100.0	99.4	99.9	100.0
	% > 95%	75.0	56.3	75.0	87.5	87.5	93.8

Std Dev = Standard deviation

MAD = Mean absolute deviation from the median

		WACSAW			Actiwatch		
WACSAW vs Actiwatch	Participant id	% correct Accuracy (adj)	% correct Sensitivity (adj)	% correct Specificity (adj)	% correct Accuracy (adj)	% correct Sensitivity (adj)	% correct Selectivity (adj)
Development cohort	1	95.8	94.3	96.7	87.8	79.6	92.4
	2	98.8	98.7	98.9	92.2	92.4	92.0
	3	99.5	99.5	99.5	88.9	96.0	85.4
	4	97.6	96.8	98.0	85.2	86.4	84.6
	5	95.8	97.2	95.2	73.8	89.7	67.2
	Mean	97.5	97.3	97.6	85.5	88.8	84.3
	Median	97.6	97.2	98.0	87.8	89.7	85.4
	Std Dev	1.7	2.0	1.7	7.0	6.2	10.2
	MAD	1.3	1.4	1.3	4.4	4.5	6.5
	min	95.8	94.3	95.2	73.8	79.6	67.2
	max	99.5	99.5	99.5	92.2	96.0	92.4
	% > 95%	100.0	80.0	100.0	0.0	20.0	0.0
Independent validation cohort	7	99.4	99.2	99.5	92.2	94.0	91.4
	9	98.8	97.1	99.6	93.9	87.6	97.5
	10	97.2	94.7	98.1	80.6	84.3	79.2
	16	98.8	97.2	99.7	89.5	88.9	89.9
	17	98.7	98.0	99.2	89.5	88.0	90.3
	18	98.8	99.3	98.5	87.6	91.6	85.3
	19	99.1	97.7	99.7	91.9	91.4	92.0
	20	98.9	97.0	100.0	89.0	72.6	98.7
	21	97.0	98.5	96.3	90.2	80.5	94.4
	22	89.9	95.5	87.3	73.4	84.2	68.2
	Mean	97.7	97.4	97.8	87.8	86.3	88.7
	Median	98.8	97.5	99.4	89.5	87.8	90.8
	Std Dev	2.8	1.5	3.9	6.2	6.3	9.2
	MAD	1.3	1.1	1.9	3.8	4.5	6.1
	min	89.9	94.7	87.3	73.4	72.6	68.2
	max	99.4	99.3	100.0	93.9	94.0	98.7
	% > 95%	90.0	90.0	90.0	0.0	0.0	20.0
Full cohort	Mean	97.6	97.4	97.7	87.0	87.1	87.2
	Median	98.8	97.2	98.9	89.0	88.0	90.3
	Std Dev	2.4	1.6	3.2	6.3	6.1	9.4
	MAD	1.4	1.2	1.8	4.1	4.6	6.6
	min	89.9	94.3	87.3	73.4	72.6	67.2
	max	99.5	99.5	100.0	93.9	96.0	98.7
	% > 95%	93.3	86.7	93.3	0.0	6.7	13.3

		WACSAW			GGIR		
WACSAW vs GGIR	Participant id	% correct Accuracy (adj)	% correct Sensitivity (adj)	% correct Specificity (adj)	% correct Accuracy (adj)	% correct Sensitivity (adj)	% correct Selectivity (adj)
Development cohort	1	95.8	94.3	96.7	88.0	78.4	92.8
	2	98.8	98.7	98.9	96.7	91.0	99.1
	3	99.5	99.5	99.5	95.9	94.2	96.7
	4	97.6	96.8	98.0	92.0	90.1	93.0
	5	95.8	97.2	95.3	86.6	90.2	85.0
	6	98.4	99.0	98.1	89.4	77.4	95.1
	Mean	97.7	97.6	97.7	91.4	86.9	93.6
	Median	98.0	97.9	98.1	90.7	90.2	94.0
	Std Dev	1.7	2.0	1.7	4.5	6.0	5.3
	MAD	1.4	1.6	1.3	3.9	3.4	3.8
	min	95.8	94.3	95.3	86.6	78.4	85.0
	max	99.5	99.5	99.5	96.7	94.2	99.1
	% > 95%	100.0	80.0	100.0	40.0	0.0	40.0
Indepentent Valication cohort	7	99.4	99.2	99.5	96.4	91.3	98.9
	8	97.7	97.3	98.0	91.2	84.6	95.1
	9	98.8	97.1	99.6	92.9	81.7	99.9
	10	97.2	94.7	98.1	92.4	85.2	94.9
	11	86.3	69.4	97.0	88.6	83.9	91.7
	12	98.9	98.4	99.1	91.0	87.6	92.8
	13	98.4	99.5	97.7	88.4	82.1	92.2
	14	96.7	99.9	95.7	91.5	93.6	90.8
	15	98.6	97.2	99.3	88.9	70.0	98.4
	16	98.8	97.2	99.7	92.7	85.6	97.3
	17	98.7	98.0	99.2	94.6	88.3	98.1
	18	98.8	99.3	98.5	91.9	88.9	93.5
	19	99.1	97.7	99.7	96.9	93.2	98.6
	20	98.9	97.0	100.0	89.5	70.1	100.0
	21	97.0	98.5	96.3	92.6	81.0	97.4
	22	89.9	95.5	87.3	85.7	86.4	85.4
	Mean	97.1	96.0	97.8	91.6	84.6	95.3
	Median	98.7	97.5	98.8	91.7	85.4	96.2
	Std Dev	3.7	7.2	3.1	3.0	6.8	4.0
	MAD	1.8	2.8	1.7	2.2	4.8	3.3
	min	86.3	69.4	87.3	85.7	70.0	85.4
	max	99.4	99.9	100.0	96.9	93.6	100.0
	% > 95%	87.5	87.5	93.8	12.5	0.0	56.3
Full cohort	Mean	97.2	96.4	97.8	91.5	85.2	94.8
	Median	98.5	97.5	98.3	91.7	86.0	95.1
	Std Dev	3.1	6.1	2.6	3.2	6.6	4.1
	MAD	1.7	2.4	1.6	2.6	5.2	3.3
	min	86.3	69.4	87.3	85.7	70.0	85.0
	max	99.5	99.9	100.0	96.9	94.2	100.0
	% > 95%	90.9	86.4	95.5	18.2	0.0	54.5

1.2	2.1	1.6	1.5	1.8	1.5
1.1	1.8	1.6	1.3	2.1	1.4
1.1	1.8	1.6	1.3	2.1	1.4
1.3	1.6	1.8	1.3	2.1	1.3
1.6	6.4	1.6	1.9	5.6	1.3
1.3	1.9	1.9	1.3	2.0	1.3
1.3	1.8	1.8	1.4	1.9	1.3
1.2	2.0	1.7	1.5	1.9	1.4
1.9	1.9	2.7	1.7	1.8	1.7
2.0	1.8	2.9	1.5	2.1	1.8
2.0	1.8	2.9	1.5	2.1	1.8
1.6	3.7	1.5	1.6	2.2	1.5
2.0	2.3	2.9	1.5	2.2	1.8
2.1	5.8	2.0	1.8	3.0	1.3
1.3	3.6	1.8	1.8	2.4	1.4
1.2	2.6	1.9	1.2	1.7	1.3
3.1	10.0	1.4	2.8	5.5	1.4
2.7	8.5	2.1	2.3	4.5	1.4
2.2	6.6	2.0	1.9	3.4	1.3
2.1	7.0	1.6	1.9	3.6	1.2