Google Research

2026-03-17

Performance analysis of updated Sleep Tracking algorithms across Google and Fitbit wearable devices

Arno Charton[1], **Linda Lei**[1], **Siddhant Swaroop**[1], **Marius Guerard**[1], **Michael Dixon**[1], **Logan Niehaus**[1], **Shao-Po Ma**[1], **Logan Schneider**[1], **Ross Wilkinson**[1], **Ryan Gillard**[1], **Conor Heneghan**[1], **Pramod Rudrapatna**[1], **Mark Malhotra**[1] **and Shwetak Patel**[1]

[1]Google Research

Background: The general public has increasingly adopted consumer wearables for sleep tracking over the past 15 years, but reports on performance versus gold standards such as polysomnogram (PSG), high quality sleep diaries and at-home portable EEG systems still show potential for improved performance. Two aspects in particular are worthy of consideration: (a) improved recognition of sleep sessions (times when a person is in bed and has attempted to sleep), and (b) improved accuracy on recognizing sleep stages relative to an accepted standard such as PSG.

Aims: This study aimed to: 1) provide an update on the methodology and performance of a system for correctly recognizing valid sleep sessions, and 2) detail an updated description of how sleep stages are calculated using accelerometer and inter-beat intervals

Methods: Novel machine learning algorithms were developed to recognize sleep sessions and sleep stages using accelerometer sensors and inter-beat intervals derived from the watch or tracker photoplethysmogram. Algorithms were developed on over 3000 nights of human-scored free-living sleep sessions from a representative population of 122 subjects, and then tested on an independent validation set of 47 users. Within sleep sessions, an algorithm was developed to recognize periods when the user was attempting to sleep (Time-Attempting-To-Sleep = TATS). For sleep stage estimation, an algorithm was trained on human expert-scored polysomnograms, and then tested on 50 withheld subject nights for its ability to recognize Wake, Light (N1/N2), Deep (N3) and REM sleep relative to expert scored labels.

Results: For sleep session estimation, the algorithm had at least 95% overlap on TATS with human consensus scoring for 94% of nights from healthy sleepers. For sleep stage estimation, comparing with the original Fitbit algorithm, Cohen's kappa for four-class determination of sleep stage increased from an average of 0.47 (std 0.14) to **0.63** (std 0.12), and average accuracy increased from 67% (std 0.10) to **77**% (std 0.078)

Conclusion: A set of new algorithms has been developed and tested on Fitbit and Pixel Watches and is capable of providing robust and accurate measurement of sleep in free-living environments.

Corresponding author(s): conorheneghan@google.com

1. Introduction

Recognized as an essential pillar of well-being, sleep's importance to health is consistently highlighted by authorities such as the American Academy of Sleep Medicine (AASM), the Centers for Disease Control and Prevention (CDC), and the American Heart Association (AHA). Individuals who monitor their sleep can gain valuable insights into their unique patterns and the various elements affecting their rest. The most accurate method for identifying sleep stages is polysomnography (PSG). This comprehensive technique involves attaching sensors to various body points to capture physiological data like brain waves (EEG), heart rhythm (ECG), blood oxygen levels (pulse oximetry), breathing patterns, leg activity, among other things. Trained specialists then visually analyze these measurements to classify sleep into five stages based on established scoring and staging rules meant to maximize inter-rater reliability: awake, REM sleep, and the three non-REM phases — N1 (light or transitional sleep), N2 (comprising most non-REM sleep), and N3 (deep or slow-wave sleep). Distinct from non-REM, REM sleep is primarily characterized by its namesake rapid eye movements, autonomic variability, and brain activity that resembles an awake state.

The Fitbit app (which is compatible with Google's Pixel Watch and Fitbit fitness tracker and smartwatch families) has provided technology for tracking sleep patterns since 2013, and has been used by millions of users, and in multiple research studies(Chinoy et al. 2021; Lee et al. 2025). The ability to estimate sleep stages was added to the product in 2017(Beattie et al. 2017). The sleep tracking ability of Fitbit devices (and similar higher end devices from other manufacturers) has been assessed to be on the same level as, if not better than, research grade actigraphy, for core sleep-wake tracking(Schyvens et al. 2025; Haghayegh et al. 2020, 2019). Some of the limitations however of wearables have been in tracking sleep of very short duration (eg, naps), and also tracking of sleep in individuals with significant nighttime movement, which can be mistaken as wakefulness. There are also reported limitations on the use of wearables to reliably track sleep in people with insomnia(de Zambotti et al. 2024).

For sleep stage estimation, the consensus is that wearables still fall short of PSG or EEG based solutions, but that progress has been made in recent years(de Zambotti et al. 2024). Accordingly, there remains ongoing interest in improving the robustness and accuracy of sleep tracking technology from wrist worn wearables. This paper outlines the methodology and performance of recent improvements to the core sleep tracking functionality used by Fitbit and Pixel Watch devices.

Fitbit sleep tracking is divided conceptually into two components: (a) firstly recognizing a period when a person is attempting to sleep, and (b) carrying out an analysis during that period of time to determine what stages of sleep were achieved.

A sleep session is a period when a person is in bed and attempts to sleep for at least 20 minutes, and can include both the main sleep (typically at night) or sleep outside of the usual sleep period (eg, naps). All of these sleep sessions contribute to a person's overall sleep profile for a day, and are important to track as reliably as possible - for example, a 1 hour nap in the middle of the day can have a significant impact on that person's "sleep pressure" later on (e.g., make it harder to fall asleep), but can also have the benefits of recovering from a prior sleep deficiency.

The sleep session classifier algorithm is based on a combination of accelerometer and optical photoplethysmogram (PPG) signals found in Pixel Watches and Fitbit trackers. The accelerometer signals contain information about body movements, and the PPG can detect individual heart beat changes, allowing the estimation of heart rate, heart rate variability, and respiration. This is described further in Section 3.1.2.

Section 3.1.3 below will detail the performance evaluation of the sleep session classifier. The assessment involved training and testing the algorithm (design and validation) with ground truth data from laboratory PSG, in-home PSG, and free living recordings.

However, simply recognizing a sleep session does not provide the complete picture of a person's sleep - it is also important to determine if the person is asleep or not and what type of sleep a person has experienced. Sleep scientists have developed rules for recognizing different "stages" of sleep, and have also discovered slightly different (but complementary) roles for each stage of sleep. For example, while all sleep stages are essential for multiple biologic functions, stage N3 ("deep" sleep) has been associated with periods of prioritized physical recovery and cellular repair.

Note that the accepted definitions of sleep stages are based primarily on signals measured from the brain (electroencephalograms—EEGs), eye movements (electrooculograms—EOGs), and somatic muscle activity (electromyograms–EMGs). Smart watches and trackers do not measure these directly, but instead measure body states from signals at the wrist, e.g., movements, heart rate, heart rate variability, estimated respiration, wrist skin temperature, etc. While there appears to be reasonably consistent mapping of sleep brain state to these variables, there will be some inter-individual variation, particularly when there is a dissociation between physiology and sleep stages (eg, people with a pacemaker). Hence, there is a theoretical upper limit on how well a wrist-worn sensor measuring cardiorespiratory and movement signals can capture sleep stages primarily defined by brain state. Furthermore, this limit on accuracy is exacerbated by inter-scorer variation on sleep staging, which results in a poor quality ground truth.

The sleep staging classifier algorithm is based on a combination of the same accelerometer and optical photoplethysmogram (PPG) signals used in the sleep session algorithm described above.

Section 3.2.1 describes the technical approaches used to develop and train the algorithm.

Section 3.2.3 details the performance evaluation of a sleep stage classifier. The assessment involved training and testing the algorithm (design and validation) with ground truth data from laboratory PSG, in-home PSG, and in-home EEG recordings.

The overall flow of sleep tracking within the Fitbit application therefore is to first examine the overall set of uploaded data at the backend to determine whether a valid sleep session has taken place - this sets the start and end times of the sleep session. The sleep stage algorithm is then applied to the data within that sleep session window to determine the hypnogram.

2. Aims

The primary aim of this paper is to explain the updated methodology for estimating sleep sessions and stages within the Fitbit application, which supports the range of Pixel Watch and Fitbit devices. It provides details about the signals used as input to the algorithms, details on the datasets used in training the machine learning models, and a summary of important aspects of the performance of these algorithms.

3. Methods

3.1. Sleep Session Estimation based on Accelerometry and Optical Photoplethysmograms

3.1.1. What is a sleep session?

Simply put, a sleep session is a period of time during which sleep occurred for a meaningful amount of time (e.g., more than twenty minutes), but which may also include times of wakefulness, before, during and after actual sleep has occurred. For example, if a person gets into bed at 11:20 PM, reads for 15 minutes, turns off the lights at 11:35 PM, then sleeps intermittently throughout the night until woken up by an alarm at 7:30 AM. At this point the user wakes up, no longer intending to sleep, and checks their phone in bed for 10 minutes, finally getting out of bed at 7:40 AM, we would define the overall sleep session as lasting from 11:35 PM to 7:30 AM. There are several key concepts that we define related to this sleep session:

1. Time in bed: this is the time that the person is physically in bed. However, it is quite common for people to spend time in bed without intending to sleep (e.g., watching TV for 30 minutes before falling asleep, reading a book for 15 minutes, etc.). We don't focus on this signal as the bed may not be the only place users sleep, and the physical location of sleeping behaviors (intentional or otherwise) is almost impossible to know with certainty.
2. Time Attempting To Sleep (TATS): this is a person's intended sleep period from the initial

attempt to sleep (e.g., by turning the lights off and closing their eyes) until they no longer plan to be sleeping (typically the final awakening, after which they begin waking activities for the day). Parameters such as sleep onset latency (SOL) are most usefully defined by considering the time between when a person has decided to fall asleep and the time when they reach their first sleep stage.

Furthermore, it is quite common for people to interrupt their sleep for various reasons during a typical night, e.g., getting up to go to the bathroom, getting up to look after a child etc. However, conceptually, if people get up for 5 minutes at night, most people do not think they have created a new “sleep session”; we extend this concept and create a new sleep session when at least 90 minutes of non-TATS is detected.

Another common form of sleep session is a daytime nap. To avoid creating false naps, we have limited this to 20 minutes or longer. However, naps are important to detect reliably as they have a significant impact on sleep pressure, overall adequacy of sleep, and circadian health.

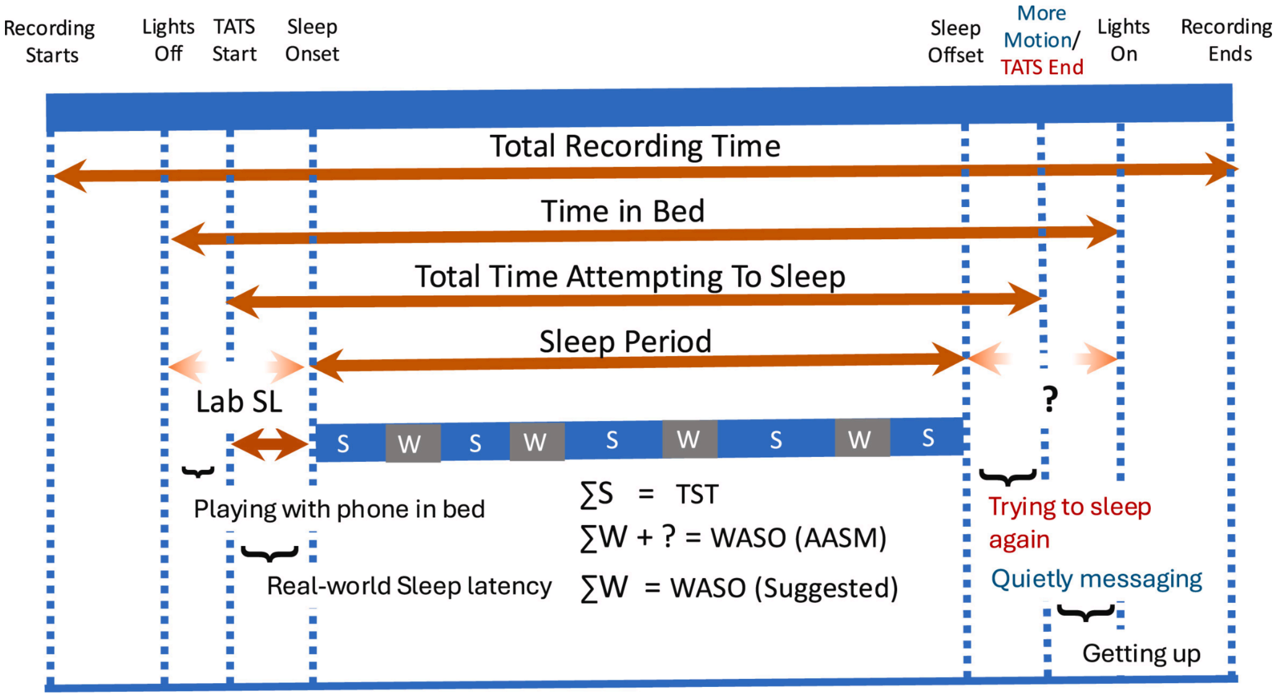


Figure 1 | We follow the definitions from the World Sleep Society(Chee et al. 2025)
Definition of various terms used to describe a sleep session. The sleep session starts at the start of the first Time Attempting To Sleep (TATS), and ends at the end of the last TATS. There may be multiple TATS in a sleep session when the user stops trying to sleep for a period of time, up to 90 minutes. A user may get in bed, play with their phone, then try falling asleep, sleep, wake up a few times, and finally wake up for the day.
Our algorithms aim to predict when TATS start, sleep session start, when the sleep session ends, when the user sleeps, and stages during their sleep.
Note that the estimates of TATS are likely to differ from truth, and in the results section we provide details on how close the start and end of estimated TATS are to ground truth.

3.1.2. Technical and Feature Description

Equipped with sensors like a 3-axis accelerometer for motion tracking (capturing both minor and significant wrist movements) and an optical photoplethysmographic (PPG) sensor using a green LED, Pixel Watches and Fitbit trackers can help ascertain a user’s sleep status. The continuous stream of data from these sensors serves as input for a specialized algorithm. This algorithm processes the signal in 30-second segments (epochs) to classify sleep. Specifically, the input to the sleep session algorithm is a combination of movement signals (both as a single 30-second summary metric, and at a 1 second resolution) and inter-beat intervals derived from the green

PPG. The algorithm also takes steps into account (as estimated from the raw accelerometer signal). [There is also an overall masking of potential sleep sessions using the "on-wrist" algorithm; i.e., no sleep sessions should be generated when the device is off-wrist.] All signals are resampled to an equivalent time-domain signal of 1 Hz for ease of computation. The algorithm is run in the cloud and currently operates on 4-hr segments of time, which are updated with a sliding window every 15 minutes. The core machine learning model applied to the 4-hr window outputs probabilities of two states: Out-of-Bed and In-Bed-Attempting to Sleep. There are a variety of post-processing and combination rules to account for the fact that the algorithm is run multiple times on overlapping data. There is also a set of heuristic rules that decide when a final decision is locked on the creation of a sleep session (e.g., during the night when a person is asleep, the sleep session is ongoing so it is premature to create the final sleep session until the person has been awake and exited bed later that morning).

3.1.3. Training and Testing

The datasets used to train, evaluate and test the sleep session algorithm were developed by using manually annotated records from internal consented volunteer users, in accordance with IRB-approved protocols. In this experiment, users wore a watch in a free-living environment and were asked to manually annotate when they went to bed, when they got up, when they attempted to sleep etc. To facilitate logging, the devices were equipped with a custom app which allowed the user to tag events such as “getting into bed”, “trying to fall asleep” etc. They were encouraged to also take naps and record these times, as well as any other unanticipated sleep events. The devices were equipped with custom software that allowed the raw accelerometer and optical photoplethysmogram data from the device to be stored and processed. To ensure the accuracy of the ground truth labeling, a visual interface was developed that allowed a panel of human reviewers to compare the user’s recollection of when events occurred versus actual timing (for example, a person might recall going to bed between 11:10PM and 11:20PM, but by examination of the raw data, that could be refined into a more precise estimate of 11:16PM). This ground truth labeling does have limitations—in general, recognizing bed-entry and bed-exit is reasonably accurate as it is accompanied by significant movement and changes in heart rate. Determining the estimated start time of when a person ’Attempts to Sleep’ is particularly challenging, as the key changes (such as reduced movement, reduction in heart rate, etc.) are subtle. Note also that the experimental data collection protocol had no mechanism for more precise measurement of bed-entry and bed-exit (e.g., video polysomnography can provide such measures, as can an under-mattress pressure pad, or bed with load cells).

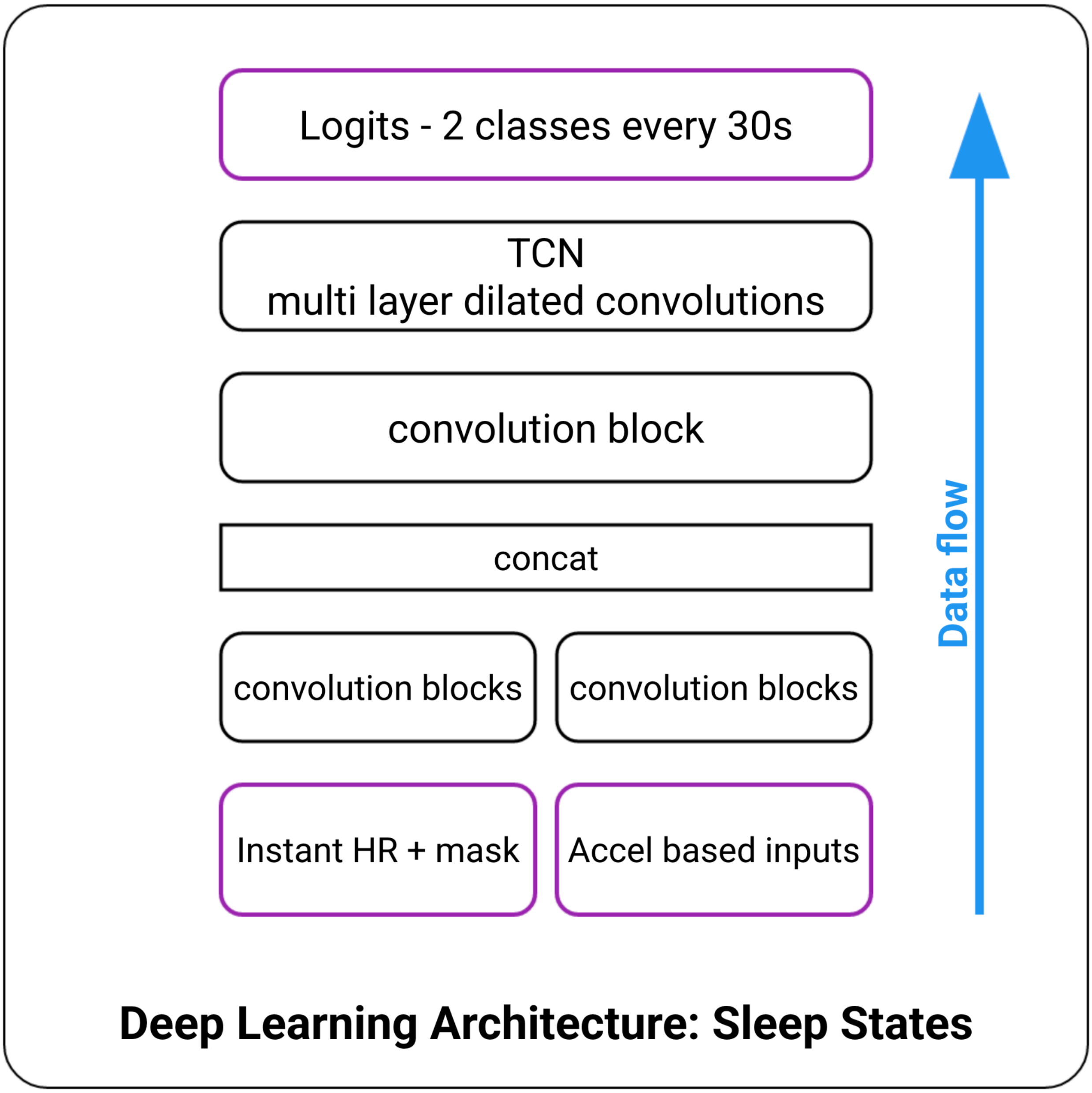


Figure 2 | A schematic representation of the algorithm used to identify sleep sessions.
A similar architecture is used for both Sleep tracking and sleep stages, with different numbers of layers. TCN=Temporal Convolutional Network. The inputs are heart rate and movement signals sampled at a 1 Hz rate. These are initially fed through convolutional blocks prior to concatenation into a single vector. That single vector is then processed through a cascade of temporal convolutional networks, with the final output layer being the raw probabilities of one of two classes at a 30-second resolution (the two classes being OutOfBed or AttemptingToSleep).

Previous evaluation of the existing sleep session algorithm in Fitbit devices had shown particular challenges in reliably identifying sleep sessions in subjects with sleep apnea. This is due to the fact that subjects with sleep apnea have large movements associated with recovery breaths, which can look similar to movements associated with wake. Likewise, there are large changes in heart rate associated with apnea events. Hence in our training and test sets we included individuals with mild/moderate/severe apnea (i.e., AHI≥ 5). The algorithms described herein are for wellness purposes only and are not intended to diagnose, treat, or monitor sleep apnea or other sleep disorders. The inclusion of apnea subjects was solely to test algorithm robustness in the presence of sleep fragmentation, not to validate apnea detection. In Section 4 on results, we provide a separate break-out of apnea and non-apnea subjects to account for that.

The performance of the automated system was then evaluated using the following key metrics. The automated algorithm estimated the start time and end time of the TATS (e.g., 11:15 PM and 6:58 AM). These were then compared with the consensus panel ground truth (GT) times (such as 11:18 PM and 6:48 AM). The percentage of sessions where the start times and end times were estimated to be within 15 minutes of the consensus GT was calculated. Furthermore, the number of sleep sessions correctly identified is an important factor. A sleep session is considered correctly identified if the overlap between the algorithm determined TATS and the ground truth is >95%. This is reported in the session coverage metric. We also track the number of "false positive" sleep sessions (e.g., where the algorithm determines a sleep session is present when in fact none occurred. This might occur, for example, if a person is sitting quietly and reading, or watching television etc.). We report this as the number of false sleep sessions/per week of continuous ground truth out of bed data. The ground truth out of bed data is derived from the same labeling process as above, where periods of high-activity sensor data without a corresponding user-annotated sleep label is labeled as "Out of Bed". This is to ensure that we're only evaluating false positives on regions of human-reviewed data determined to be non-sleeping. The available labels and sensor data were split into a training and validation set with representative cohorts for age, sex and BMI.

Table 1 | Subject demographics for sleep session training and validation

	Training Set	Validation Set
Subjects	122	47
Nights of data	3530	1596
Age Brackets	[18-29 yrs]:16 (13%) [30-49 yrs]:71 (58%) [≥ 50 yrs]:35 (29%)	[18-29 yrs]:6 (13%) [30-49 yrs]:22 (47%) [≥ 50 yrs]:19 (40%)
Sex	Female: 37 (30%) Male: 80 (70%) Transgender/Non Binary/Decline to State: 0	Female: 15 (32%) Male: 32 (68%) Transgender/Non Binary/Decline to State: 0
BMI Brackets	≤ 18.5kg/m^2: 2 (2%) 18.5-30kg/m^2: 70 (57%) ≥ 30kg/m^2: 48 (41%)	≤ 18.5kg/m^2: 0 (0%) 18.5-30kg/m^2: 29 (62%) ≥ 30kg/m^2: 18 (38%)
Sleep Duration (mins)	447±161	474±176

3.2. Sleep Staging based on Accelerometry and Optical Photoplethysmograms

3.2.1. Technical and Feature Description

As noted above, the trackers and smartwatches are equipped with a 3-axis accelerometer for motion tracking (capturing both minor and significant wrist movements) and a set of green, red, and infrared optical plethysmography sensors. An on-device algorithm extracts cardiac interbeat-intervals and movement coefficients (compressed representation of accel state during the measurement period) from the raw data, and transmits these to the back-end Fitbit server for processing. The sleep stage processing is triggered once a sleep session has been finalized. The sleep staging algorithm processes the signal in 30-second segments (epochs) to classify sleep into four categories: Wake, Light, Deep, or REM. This classification system mirrors established human scoring practices for sleep analysis, which are explained below. The algorithm's training and testing relied on a robust collection of overnight polysomnography (PSG) recordings, gathered both in laboratory settings and at users' homes. Each 30-second epoch within these PSG recordings was assigned a definitive ground truth stage label by expert human PSG technologists. The raw data was collected from participants who provided informed consent for research purposes in a number of IRB approved studies from a range of adult subjects (sampled to uniformly cover decade-based age brackets and genders) who were both free from sleep disorders as well as having various sleep-related pathologies (eg, varying severities of obstructive sleep apnea, insomnia disorder, and periodic limb movements of sleep). In terms of output, the 'Wake' and 'REM' labels were meant to directly correspond to the clinically recognized stages of the same names. The 'Deep' sleep label corresponds to what is clinically known as N3 (non-REM stage 3), and the 'Light' sleep label groups the N1 and N2 clinical stages together, noting that N1 usually forms only a small part of a night's sleep.

One of the challenges in sleep staging is establishing the ground truth due to both inter-scorer reliability, and also inherent challenges in assigning arbitrary definitions to inherently analog continuous biological processes. In particular, we have found many cases where the scoring of N3 sleep differs between the wearable-based estimate and the underlying EEG scoring. Upon further investigation, we observed that in many of these cases, the spectral analysis of the EEG showed strong evidence of delta wave activity but was not scored as N3 due to not achieving the 75 μV threshold. This is a known deficiency in how N3 sleep is manually scored by humans applying the AASM's criteria, resulting in an unreliable, nonphysiologic ground truth label for model development(Younes et al. 2018). Therefore, we re-scored both the training and test data to reflect more accurately the underlying physiology. We adapted the methodology used to derive Slow Wave Power as outlined in(Davidson et al. 2025). The net effect of this re-scoring was to increase the average percentage of deep sleep from 10.3% to a more physiologically plausible and epidemiologically appropriate 17.3% which is a more realistic distribution of expected sleep for this population(Boulos et al. 2019).

The system was evaluated on an internal set of records obtained under IRB approval from adults with corresponding scored PSG sleep stage labels. The system was tested using a withheld data set of 50 records, containing subjects with sleep apnea (12 healthy patients, 16 with mild OSA, 22 with moderate and severe OSA). The overall performance of the system was evaluated by calculating two stages (wake versus sleep) and four stage accuracy and Cohen's kappa values. We also considered the resulting distributions of stage durations at a population level, and considered the goodness-of-fit for individual sleep stage durations on a per-night basis Details on the demographics of the training and validation data set are given in Table 1

3.2.2. Training and Testing

The sleep staging algorithm was developed by initially training on a set of PSG data representative of expected users, followed by independent testing on a set of nightly records it used in the training process. Table 2 gives the details on the data sets used for training and testing.

Table 2 | Subject demographics for sleep stage training and validation

	Validation Set
Subjects	50
Nights of data	50
Age Brackets	[18-29 yrs]:7 (14%) [30-49 yrs]:24 (48%) [≥ 50 yrs]:19 (38%)

Sex	Female: 26 (52%) Male: 24 (48%) Transgender/Non Binary/Decline to State: 0
BMI Brackets	≤ 18.5kg/m^2: 12 (24%) 18.5-30kg/m^2: 13 (26%) ≥ 30kg/m^2: 25 (50%)
Total Sleep Time (mins)	474±176

3.2.3. Four-stage classification performance

In a multi-class problem, where you wish to compare the labeling from two different observers, a confusion matrix is a standard method for presenting results. It juxtaposes the actual ground truth labels (represented in rows) against the algorithm's predictions (in columns). This format clearly displays classification accuracy in the diagonal cells and the specific types and proportions of errors in the off-diagonal cells. Each value within the matrix indicates the percentage of true epochs for a specific state (the row) that the algorithm classified into each of the four possible states (the columns). In the Results section, we will provide the confusion matrix for the test set. Confusion matrices and metrics are computed per night, then averaged over the dataset. Another way to summarize performance accuracy is through Cohen's kappa coefficient (κ), a measure of performance accuracy that takes into account the role of chance.

It is defined as

$$\kappa = \frac{\%Observed\ Agreements - \%Agreements\ by\ chance}{1 - \%Agreements\ by\ chance}$$

This is a measure of inter-rater agreement, where the two raters in our case are the expert sleep technician(s) (who scored the polysomnogram recordings) and the automated sleep staging system. κ is a chance-adjusted measure of agreement which ranges from -1 to 1, where 1 indicates perfect agreement. A κ of 0 indicates the agreement between the raters is equivalent to chance. Cohen proposed that κ be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. A weighted kappa could also be useful, as in practice not all misclassifications are of equal importance (e.g. mislabeling Deep sleep as Wake is a 'worse' error than labeling a Light sleep epoch as REM). However other researchers in this field have tended to provide unweighted kappa scores, so for ease of comparison that is what we report. As well as kappa, we also report on overall accuracy (performance) which is the percentage of epochs correctly labeled relative to the gold standard, as well as class-specific sensitivity and specificity, and level of agreement for ease of comparison with other work. We also provide a two-stage classification of sleep/wake, where the sleep sensitivity is defined as the percentage of true sleep epochs correctly labeled as sleep (combining Light, Deep, and REM classes). The wake sensitivity is defined as the percentage of true wake epochs correctly labeled as wake. As well as per-epoch accuracy, overall durations are of importance in assessing the quality of sleep, so we also provide results on the estimated durations of sleep in the various labeled stages. A further analysis of interest is to determine whether there is a systematic bias in either under-estimating or over-estimating the durations in the various sleep stages(Menghini et al. 2021).

4. Results

The algorithms were trained on representative data as described above, and then tested on a validation set that was not used in the training of the system.

4.1. Results on Sleep Session Recognition

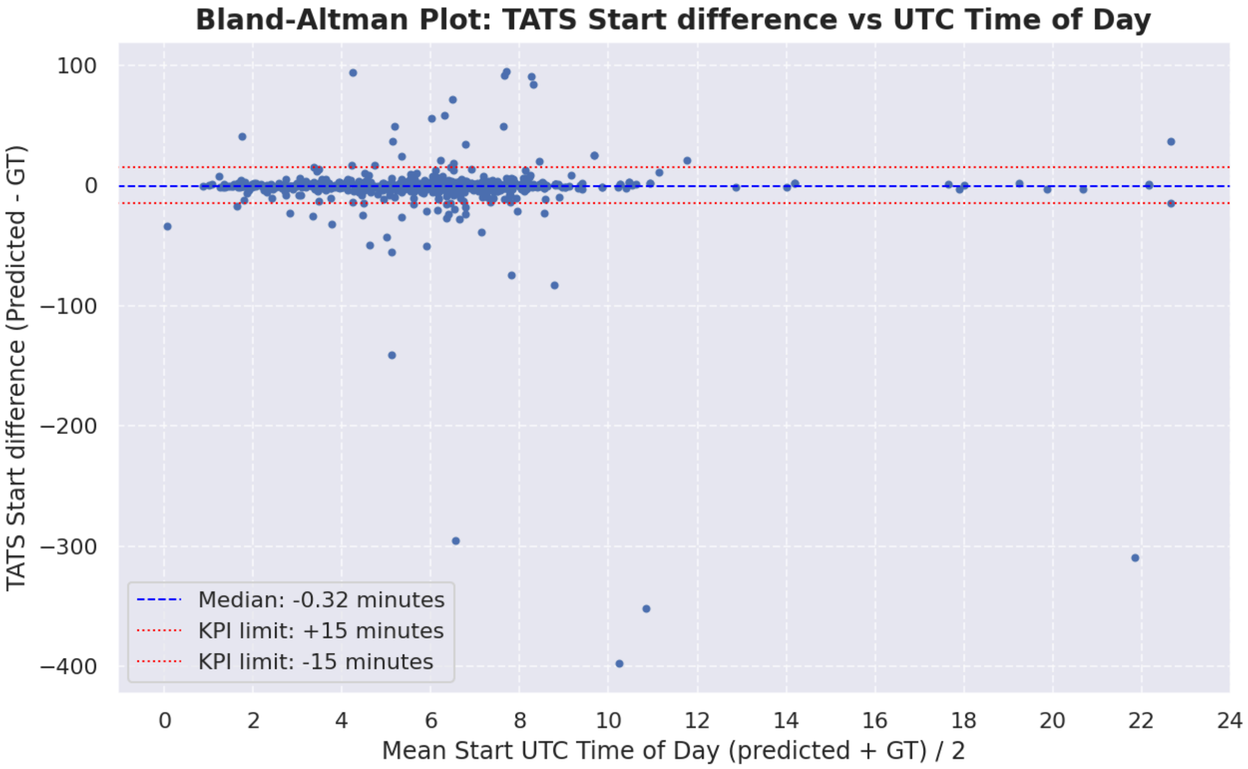
To give a reasonable estimate of expected performance, we used the following metrics.

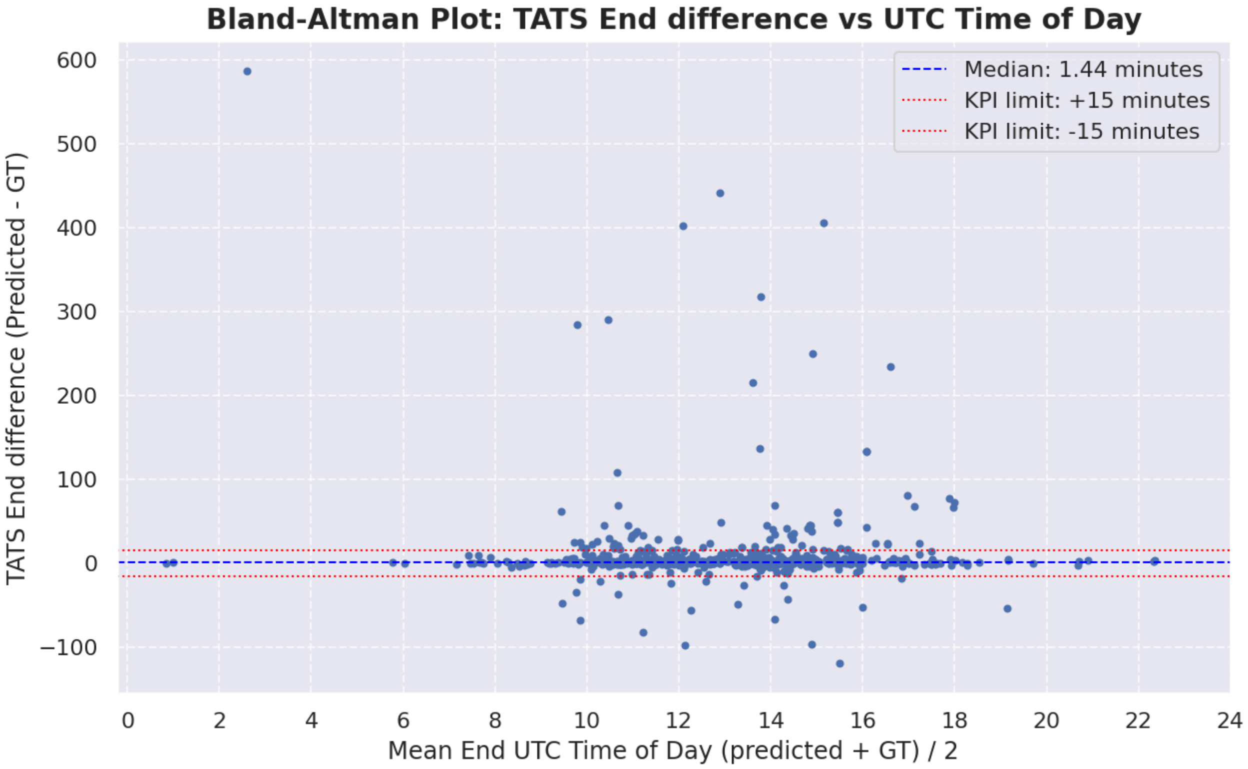
1. TATS-Start15—Percentage of sleep sessions where the start of TATS was within ± 15 minutes of expert denoted TATS-start
2. TATS-End15—Percentage of sleep sessions where the end of TATS was within ± 15 minutes of expert denoted TATS-end
3. TATS95—Percentage of sleep sessions where the ground truth was at least 95% covered by the estimated TATS.
4. False Positive—Percentage of unmatched TATS that overlap in any way with a section of time that experts denoted contained no sleep.
5. False Negative—Percentage of experts denoted TATS that do not have any overlapping algorithm-generated TATS.

Table 3 | Results of sleep session tracking algorithm

Validation Set	Non-Apnea		Apnea	
	Original	**New model**	Original	**New model**
TATS-Start15	79.3%	**91.4%**	67.1%	**81.6%**
TATS-End15	61.3%	**83.6%**	64.7%	**85.2%**
TATS95	85.8%	**93.7%**	50.2%	**76.2%**
False Positive	2.1%	**0.6%**	5.3	**1.6%**
False Negative	3.5%	**0.6%**	3.4%	**1.6%**

Figure 3 | Sleep session start and end latency by ground truth sleep session duration, validation set, non-apnea dataset. The dotted red lines show the plus and minus 15 minutes limits for the results above.





From the table of results, we note that typically 80-90% of sleep sessions correctly identify the start of the TATS period within 15 minutes of annotated ground truth. To put these results in context, recall that these metrics TATS have no exact ground truth (there is no video polysomnography, pressure sensitive mattress, etc. which can provide more precise metrics of TATS). Rather, the ground truth for TATS is derived using a panel of up to eight human reviewers looking at the raw accelerometer, heart rate signals, steps, light sensor, and user annotations. In fact, the typical agreement level for a single human scorer on the same data is typically 80-90%. By its nature, TATS is a psychological construct (even with their eyes closed in bed, a person may not intend to fall asleep but may instead be meditating or relaxing).

Of all the parameters, recognition of the start time of TATS is most robust with an average performance of about 91%, perhaps reflecting the fact when people often first start trying to sleep, their movement and heart rate often have a distinctive signature reflective of their behavioral intent.

4.2. Results on Sleep Stage Recognition

Table4 shows the sleep stage performance improvement over time, the table compares the four stages Cohens-Kappa and accuracy of previous Fitbit releases. Figure 4 shows the averaged confusion matrix for the test set, N=50. The table also includes the true-positive and false-negative values for each class. For example, on average, 69% of Wake epochs are correctly labeled as Wake, and 3% of Wake epochs are incorrectly labeled as REM. The highest error rates occur with Deep/Light mislabeling (often with N3 sleep being mislabeled as Light), and incorrect labeling of Wake states as Light sleep. The average four-stage kappa value in the validation set was 0.63 (SD 0.12). Figure 5 shows the reduced-complexity sleep staging where all the sleep-stages have been collapsed into a single class. As an alternative interpretation of the model performance, it is instructive to view the estimated sleep duration in each stage for each night, to see if performance is robust across subject-nights. Figure 6 provides a scatter plot for each subject-night of the estimated duration in each stage versus the gold-standard duration. Figure 7 shows the total duration per stage Bland-Altman plots. Figure 8 shows the per stage distribution of durations.

Table 4 | Performance results comparison from previous Fitbit releases on the same Test set.
A jump in the sleep stage performance was released in July 2025, the new model improves the performance further.

	Cohens Kappa 4		Accuracy 4	
Model	Average	Std	Average	Std
New model	**0.63**	**0.12**	**0.77**	**0.08**
July 2025 model	0.56	0.13	0.71	0.10
Original	0.47	0.14	0.67	0.10

Figure 4 | Confusion matrix for Four-Class Sleep Staging. Shows the average confusion matrix and std (computed by averaging the confusion matrices for each session from the test set, N=50). The labels are Awake, Light (N1/N2), Deep (N3), REM.
Performance metrics shared in this document were computed per night and averaged, this gives equal weight to each night, short or long.

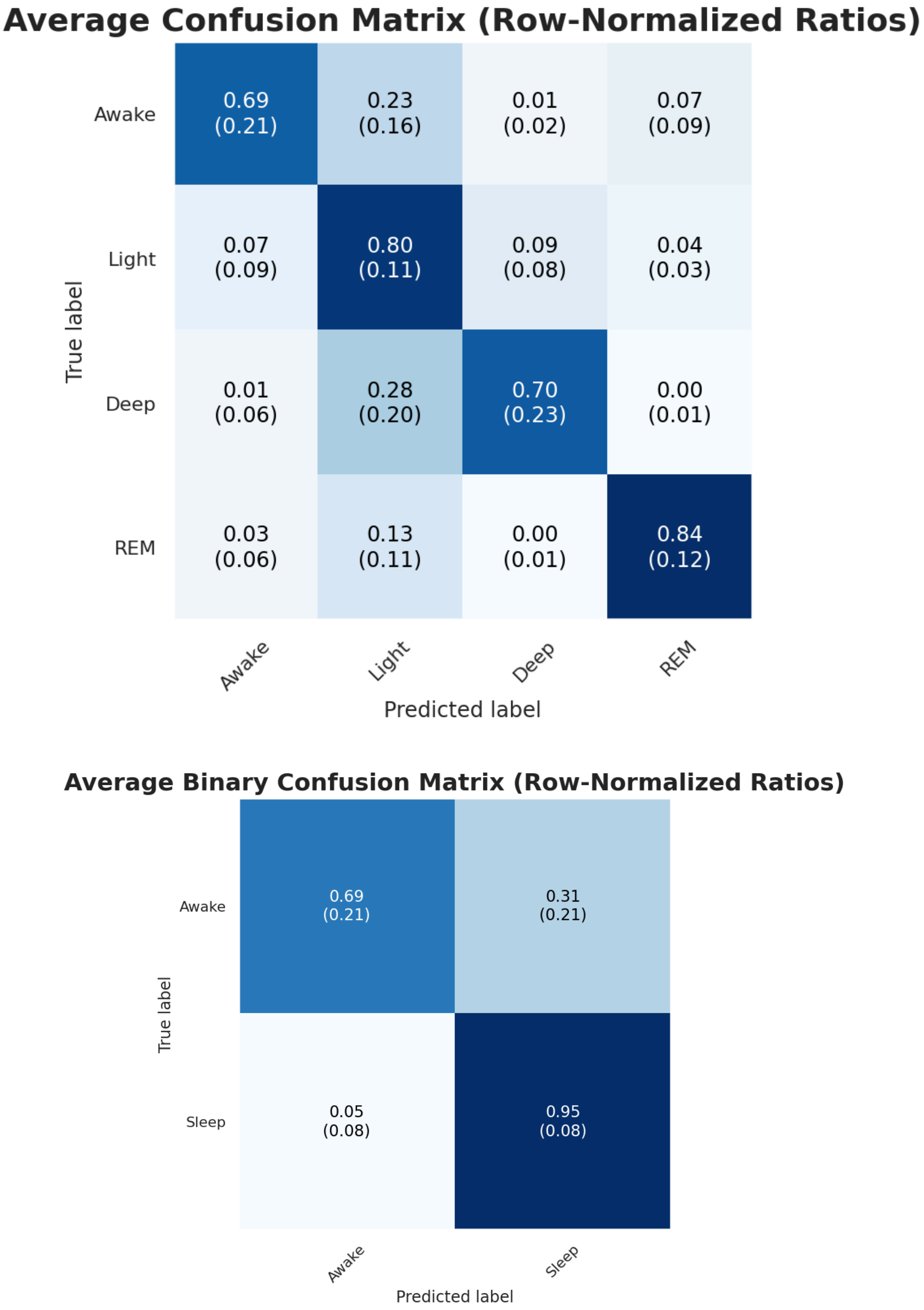


Figure 5 | Confusion matrix for Awake/Sleep by combining sleep stages Light (N1/N2), Deep (N3), and REM into the Sleep class. Shows the average and std.

An important aspect of the classifier and post-processing system is to minimize any potential bias towards a particular sleep stage. For example, if we would like to use this system to estimate population level means of the distribution of times in the various stages, it is desirable that the classifier has minimal bias. Figure 8 shows a boxplot for the duration in each sleep state across the 50 participants in the test set. The boxplot shows the median, interquartile ranges, and maximum/minimum of the data which falls within a distance less than 1.5 times the interquartile range (at the ends of the whiskers), as well as outliers. It can be seen that the model does well in adhering closely to population means by stage.

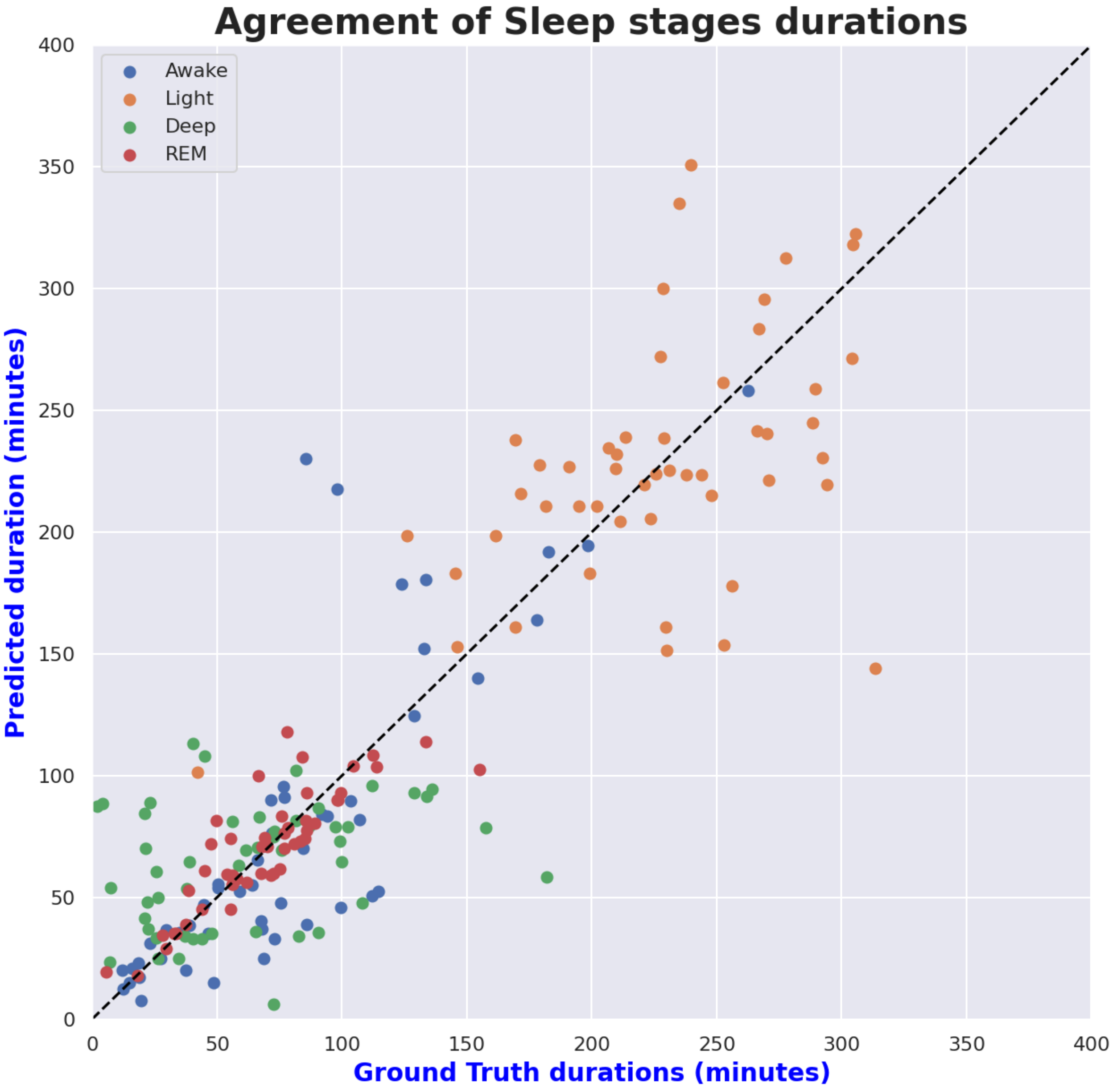


Figure 6 | Predicted versus actual time spent in each sleep stage, for all 50 subject-nights in the test set. The diagonal dotted line represents perfect agreement between the model's estimated stage duration and the ground truth duration. The goodness-of-fit parameter $R^2$is 0.672

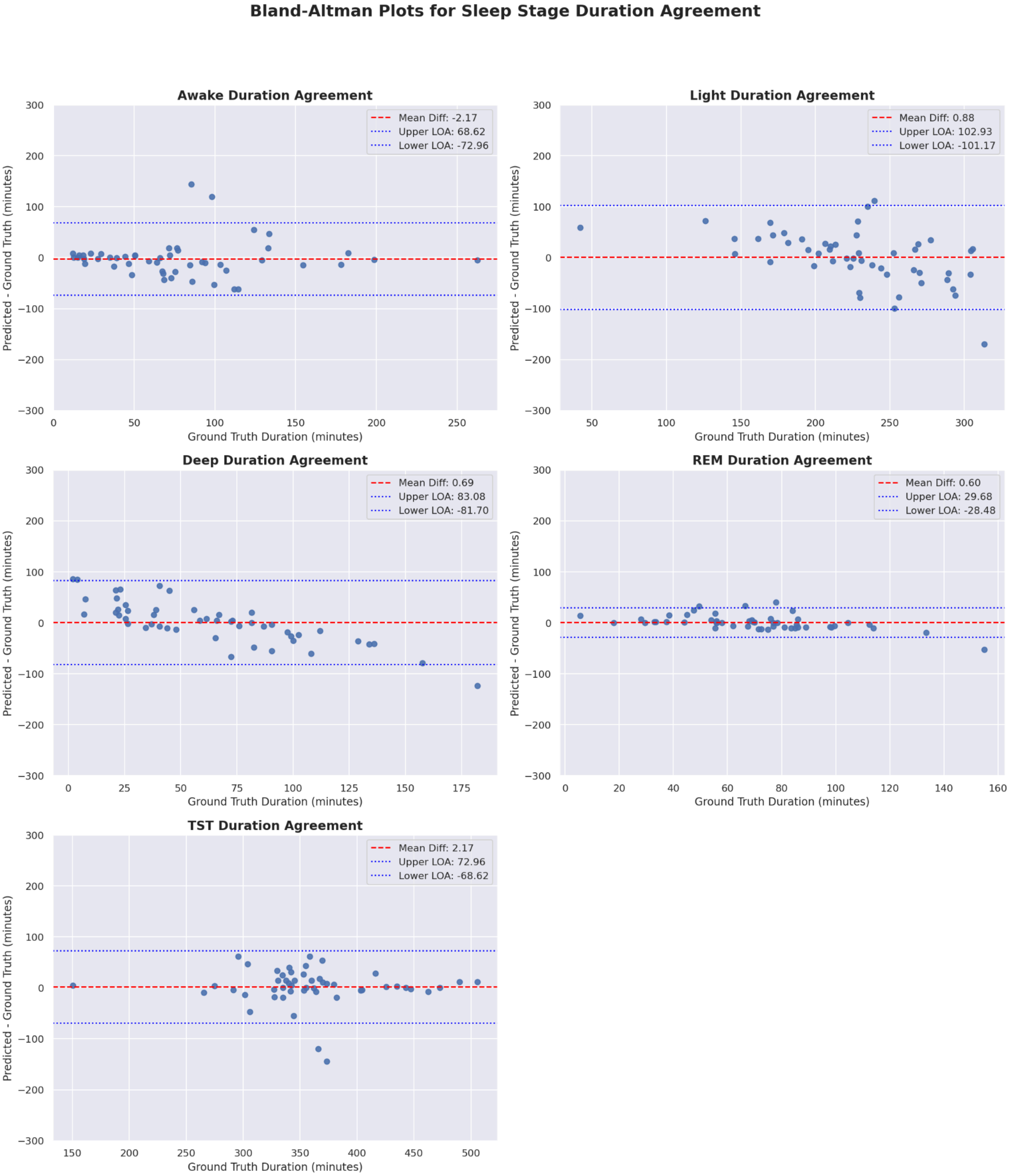


Figure 7 | Per stage total duration Bland-Altman plots. Note the operating point of the model was calibrated to reduce bias. For the Deep (N3) stage, we have a negative correlation between the difference of predicted minus ground truth, and the ground truth duration, the linear regression slope is -0.822, and the intercept is 52.7 minutes.

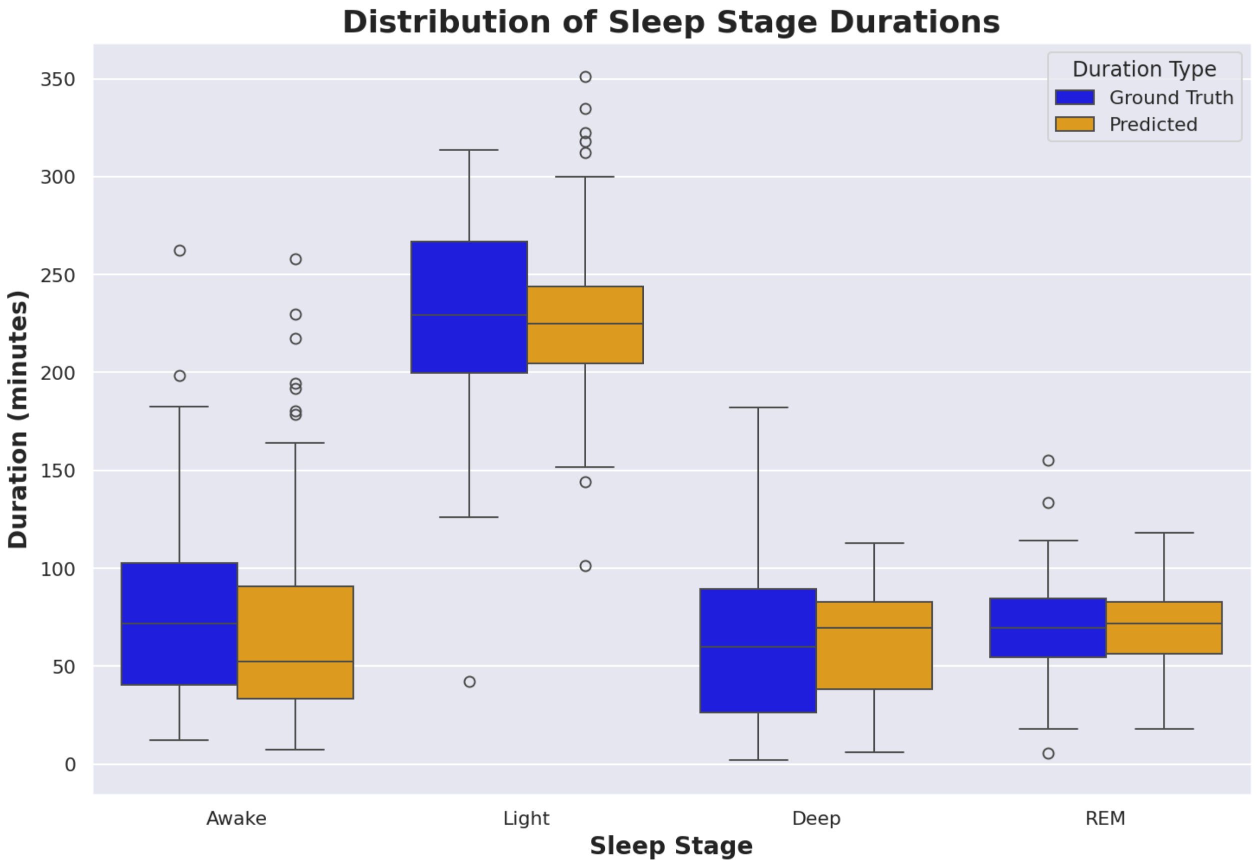


Figure 8 | This shows the population-level (N=50) distribution of sleep stages estimated by the mode, as compared to the ground truth from the PSG scoring in the performance validation hold-out set.

To understand potential confounders impacting the multistage kappa values, additional analyses were conducted by demographics and sleep abnormalities, measured by the gold standard scoring. When modeled together, the normalized confounder coefficients in a general linear model provide insight into the relative impact of confounders. Higher stage transitions and higher AHI impacted kappa performance, which was not unexpected because factors that directly influence sleep architecture — such as frequent stage transitions — are predicted to be more challenging for classification. Figure 9 shows the impact of apnea severity on kappa.

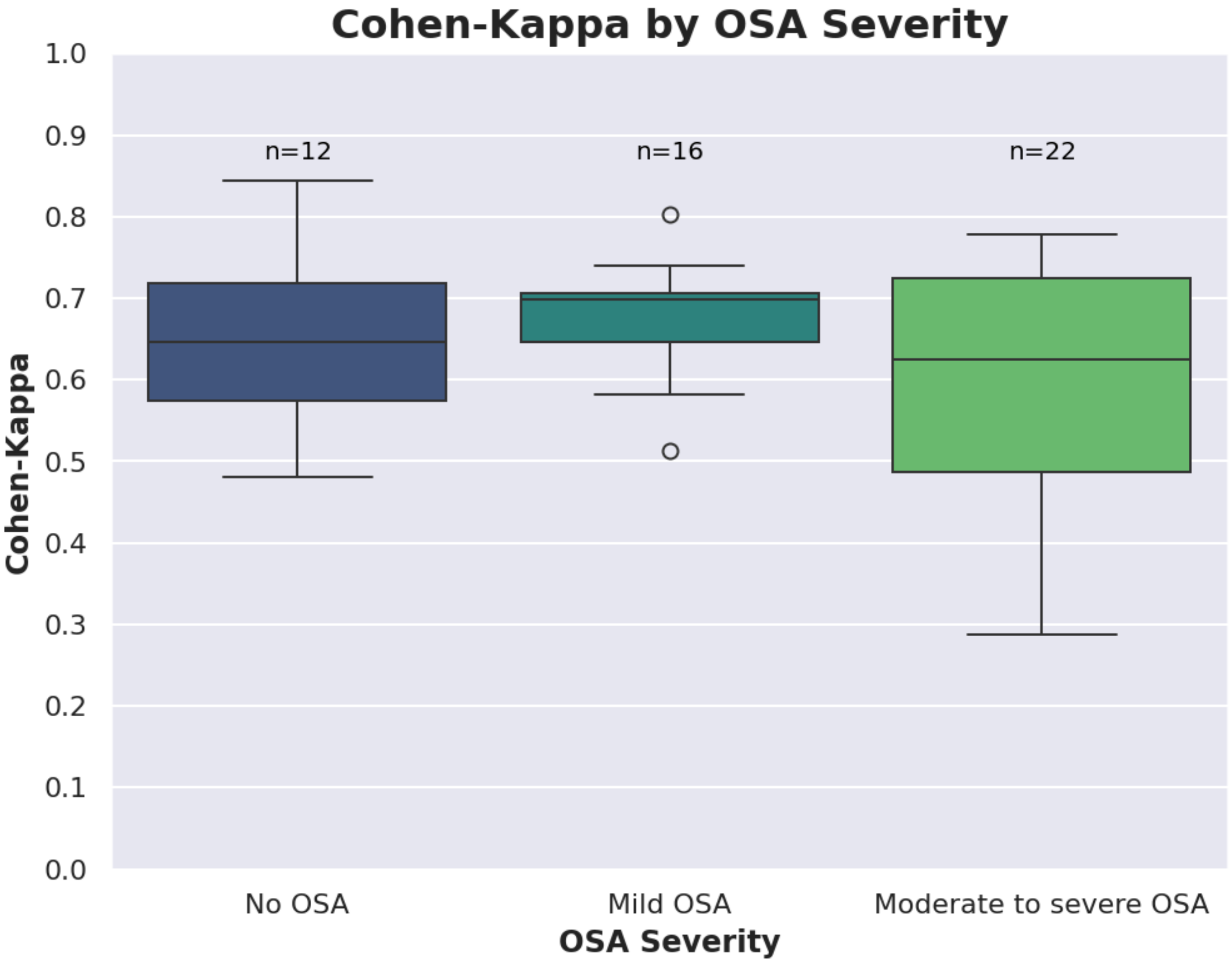


Figure 9 | The distribution of observed Cohen's kappa values when bracketed by OSA severity (Obstructive Sleep Apnea). It is typically harder to reliably stage sleep data from subjects with higher AHI. The No OSA category contains AHI of 5 or below, Mild is 5 < AHI < 15, and moderate to severe has AHI >= 15.

5. Discussion and Conclusions

This study describes the training and validation of a set of updated algorithms for use with Fitbit and Pixel trackers and watches. The training and validation process included a wide range of ages and BMIs, as well as a significant number of subjects with sleep apnea, which is a key confounding factor for accurate sleep staging. The algorithm introduces the concept of estimating time attempting to sleep (TATS) which has not previously been reported in consumer sleep tracking wearables, but which has been suggested in the Consumer Technology Association's standard for consumer sleep wearables(*Definitions and Characteristics for Wearable Sleep Monitors* 2022) and the World Sleep Society's recommendations for the use of wearable consumer health trackers that monitor sleep(Chee et al. 2025).

There have not been many studies on the ability of consumer wearables to track TATS, so it is hard to provide any comparative performance figures.

There have been several studies on the accuracy of consumer sleep wearables for sleep staging. In our prior evaluation of the initial Fitbit sleep staging algorithm, the observed kappa was 0.52(Beattie et al. 2017), so this new updated algorithm provides a meaningful boost in sleep staging performance from 0.52->0.63. In an independent recent study comparing consumer wearable devices for sleep staging, researchers observed kappas from 0.3-0.44 across a range of commercially available devices in a population of between 20-50 subjects for each device(Lee et al. 2023), well below the observed performance of this updated algorithm. However, head-to-head device performance needs to be assessed in future studies to make meaningful comparisons. For additional context, there appears to be an upper limit of performance, as assessed by Cohen's kappa, based on the inherent limits of inter-scorer reliability of manually scored sleep studies, which are approximately 0.7 to 0.8(Lee et al. 2022).

The results discussed here have included both nominally healthy sleepers and people with varying severities of sleep apnea. In future work, we will provide results on another significant set of sleepers, those people reporting symptoms of insomnia.

This paper provides a summary of a set of updated algorithms which are implemented for the set of Fitbit and Pixel watches in early 2026. It highlights improved performance in the detection of sleep sessions which will lead likely to more robust estimates of sleep-wake patterns. It also demonstrates improved performance in sleep staging - we believe that increased accuracy and robustness of sleep stage estimation will prove important for deriving longer-term sleep health insight and knowledge for users of consumer health trackers for personal or research purposes(Zheng et al. 2024). For example, there is growing interest in whether longitudinal changes in sleep architecture may be indicators of impending cognitive decline(Kim 2024). However, since currently such studies have relied on lab-based polysomnograms, they have been limited in scale and duration - reliable sleep stage estimation with consumer wearables will allow such population-level hypotheses to be better investigated, as has already been demonstrated in the All of Us research study(Zheng et al. 2024).

References

Beattie, Z., Y. Oyang, A. Statan, et al. 2017. "Estimation of Sleep Stages in a Healthy Adult Population from Optical Plethysmography and Accelerometer Signals." *Physiological Measurement* 38 (11): 1968–1979.

Boulos, Mark I., Trevor Jairam, Tetyana Kendzerska, James Im, Anastasia Mekhael, and Brian J. Murray. 2019. "Normal Polysomnography Parameters in Healthy Adults: A Systematic Review and Meta-Analysis." *The Lancet. Respiratory Medicine* 7 (6): 533–543.

Chee, Michael WI, Mathias Baumert, Hannah Scott, et al. 2025. "World Sleep Society Recommendations for the Use of Wearable Consumer Health Trackers That Monitor Sleep." *Sleep Medicine* 131 (106506): 106506.

Chinoy, Evan D., Joseph A. Cuellar, Kirbie E. Huwa, et al. 2021. "Performance of Seven Consumer Sleep-Tracking Devices Compared with Polysomnography." *Sleep* 44 (5). https://doi.org/10.1093/sleep/zsaa291.

Davidson, Shaun, Rachel Sharman, Simon D. Kyle, and Lionel Tarassenko. 2025. "Is It Time to Revisit the Scoring of Slow Wave (N3) Sleep?" *Sleep* 48 (10). https://doi.org/10.1093/sleep/zsaf063.

Definitions and Characteristics for Wearable Sleep Monitors. 2022. Consumer Technology Association. https://www.thensf.org/wp-content/uploads/2022/10/ANSI-CTA-NSF-2052.1-A-FINAL.pdf.

Haghayegh, Shahab, Sepideh Khoshnevis, Michael H. Smolensky, Kenneth R. Diller, and Richard J. Castriotta. 2019. "Accuracy of Wristband Fitbit Models in Assessing Sleep: Systematic Review and Meta-Analysis." *Journal of Medical Internet Research* 21 (11): e16273.

Haghayegh, Shahab, Sepideh Khoshnevis, Michael H. Smolensky, Kenneth R. Diller, and Richard J. Castriotta. 2020. "Performance Assessment of New-Generation Fitbit Technology in Deriving Sleep Parameters and Stages." *Chronobiology International* 37 (1): 47–59.

Kim, Kwang-Youn A. 2024. "What the Changes in Sleep Architecture Tell You about Cognitive Decline-an Editorial." *Sleep* 47 (10): zsae180.

Lee, Taeyoung, Younghoon Cho, Kwang Su Cha, et al. 2023. “Accuracy of 11 Wearable, Nearable, and Airable Consumer Sleep Trackers: Prospective Multicenter Validation Study.” *JMIR mHealth and uHealth* 11 (November): e50983.

Lee, Young Jeong, Jae Yong Lee, Jae Hoon Cho, Yun Jin Kang, and Ji Ho Choi. 2025. “Performance of Consumer Wrist-Worn Sleep Tracking Devices Compared to Polysomnography: A Meta-Analysis.” *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine* 21 (3): 573–582.

Lee, Yun Ji, Jae Yong Lee, Jae Hoon Cho, and Ji Ho Choi. 2022. “Interrater Reliability of Sleep Stage Scoring: A Meta-Analysis.” *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine* 18 (1): 193–202.

Menghini, Luca, Nicola Cellini, Aimee Goldstone, Fiona C. Baker, and Massimiliano de Zambotti. 2021. “A Standardized Framework for Testing the Performance of Sleep-Tracking Technology: Step-by-Step Guidelines and Open-Source Code.” *Sleep* 44 (2). https://doi.org/10.1093/sleep/zsaa170.

Schyvens, An-Marie, Brent Peters, Nina Catharina Van Oost, et al. 2025. “A Performance Validation of Six Commercial Wrist-Worn Wearable Sleep-Tracking Devices for Sleep Stage Scoring Compared to Polysomnography.” *Sleep Advances: A Journal of the Sleep Research Society* 6 (2): zpaf021.

Younes, Magdy, Samuel T. Kuna, Allan I. Pack, et al. 2018. “Reliability of the American Academy of Sleep Medicine Rules for Assessing Sleep Depth in Clinical Practice.” *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine* 14 (2): 205–213.

Zambotti, Massimiliano de, Cathy Goldstein, Jesse Cook, et al. 2024. “State of the Science and Recommendations for Using Wearable Technology in Sleep and Circadian Research.” *Sleep* 47 (4). https://doi.org/10.1093/sleep/zsad325.

Zheng, Neil S., Jeffrey Annis, Hiral Master, et al. 2024. “Sleep Patterns and Risk of Chronic Disease as Measured by Long-Term Monitoring with Commercial Wearable Devices in the All of Us Research Program.” *Nature Medicine* 30 (9): 2648–2656.