Estimating Breathing Disturbances and Sleep Apnea Risk from Apple Watch

September 2024

Overview

Users can measure Breathing Disturbances during sleep tracking with Apple Watch Series 9 or later and Apple Watch Ultra 2 (excluding Apple Watch SE) to better understand interruptions in their respiratory patterns. Apple Watch also uses this data in an algorithm that detects signs of moderate to severe sleep apnea.

Sleep apnea is a common and treatable disorder that negatively impacts people's health and quality of life. Despite advances in public awareness about the importance of sleep, most cases of sleep apnea remain undiagnosed. Apple Watch can track Breathing Disturbances during sleep and provide notifications of possible sleep apnea if Breathing Disturbances values reach a level associated with moderate to severe sleep apnea.

This article describes the development and validation of the Sleep Apnea Notification Feature. Apple developed the Apple Watch algorithm, which uses accelerometer data to estimate Breathing Disturbances, with a large and diverse training set of adults undergoing simultaneous recordings from Apple Watch and a ground truth reference. This included either in-lab polysomnography (PSG) or home sleep apnea test (HSAT) kits, which measure apnea and hypopnea events — the foundation of the apnea-hypopnea index (AHI).

Introduction

Sleep apnea is a treatable sleep disorder characterized by repeated disturbances in breathing during sleep. Untreated sleep apnea can increase the risk of health issues such as high blood pressure, heart attack, and stroke. The health implications of sleep apnea are most significant for moderate to severe disease, defined by an apnea-hypopnea index (AHI) — a measure of the average number of apnea or hypopnea events per hour over a recording interval — of 15 or more. Clinical guidelines indicate that moderate to severe sleep apnea warrants treatment regardless of symptoms and that treatment for mild sleep apnea is recommended if symptomatic (Epstein et al. 2009). Fortunately, well-established treatment options are available for this condition, such as positive airway pressure devices (for example, CPAP), dental appliances, and surgical options.

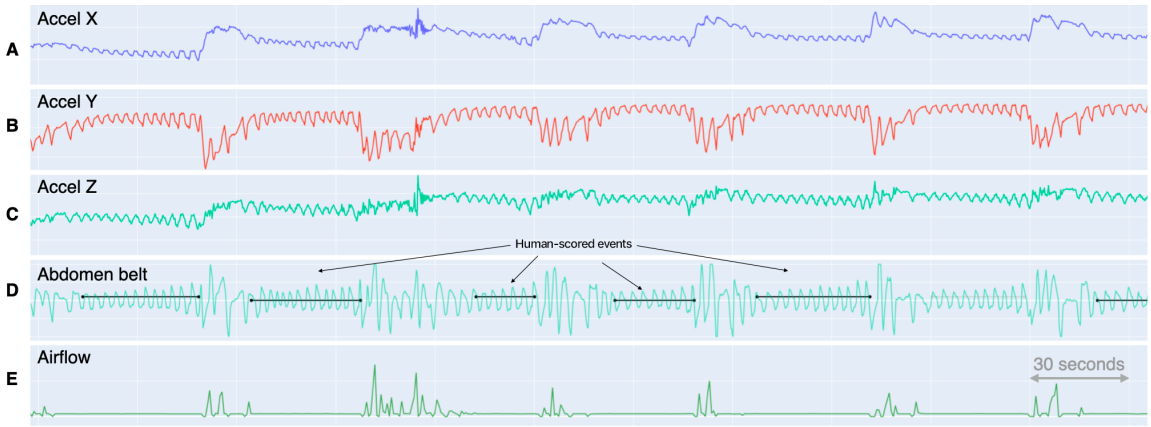
The prevalence of moderate to severe sleep apnea is estimated at approximately 10% of the global adult population, based on a strict scoring convention known as the 4% hypopnea rule (Benjafield et al. 2019). Research also suggests that up to 80% of sleep apnea cases remain undiagnosed (Peppard et al. 2013). While some people experience symptoms such as daytime drowsiness or gasping for breath at night, other people may not have any symptoms.

Accelerometer-Based Breathing Analysis

Technical and Feature Description

Apple Watch tracks movement with triaxial accelerometer signals, which capture coarse motion of the body as well as fine movements including motion associated with breathing. Apple developed an algorithm that uses the accelerometer time series data to classify Breathing Disturbances that occur during sleep tracking, which are temporary interruptions in the breathing pattern. Figure 1 demonstrates the relationship of the Apple Watch accelerometer data (X, Y, and Z) alongside more traditional methods of identifying disruptions to breathing, including abdomen belts and airflow meters.

The Sleep Apnea Notification Feature contains two components: a nightly Breathing Disturbances measurement and a notification sent to the user if the Breathing Disturbances values are elevated over a 30-day period, indicating signs of potential moderate to severe sleep apnea. Note that neither component is intended or cleared for use by people who are already diagnosed with sleep apnea. Breathing Disturbances aren't equivalent to the AHI.



Human-scored events indicate interruptions in the respiratory pattern alternating with recovery breaths.

Figure 1. Apple Watch Accelerometer Signals and Ground Truth Respiration Signals

Initial use of this feature requires a short onboarding process and sleep tracking to be enabled on Apple Watch. Once active, Breathing Disturbances values can be viewed in the Health app. Tracking Breathing Disturbances helps users understand their sleep and the behaviors or lifestyle factors that may affect their Breathing Disturbances levels. The notification feature alerts users who may be at risk of moderate to severe sleep apnea. An example of Breathing Disturbance data viewed in the iPhone Health app is shown in Figure 2. In this example, the majority of nights were classified as not elevated, so the displayed period is summarized as Not Elevated.

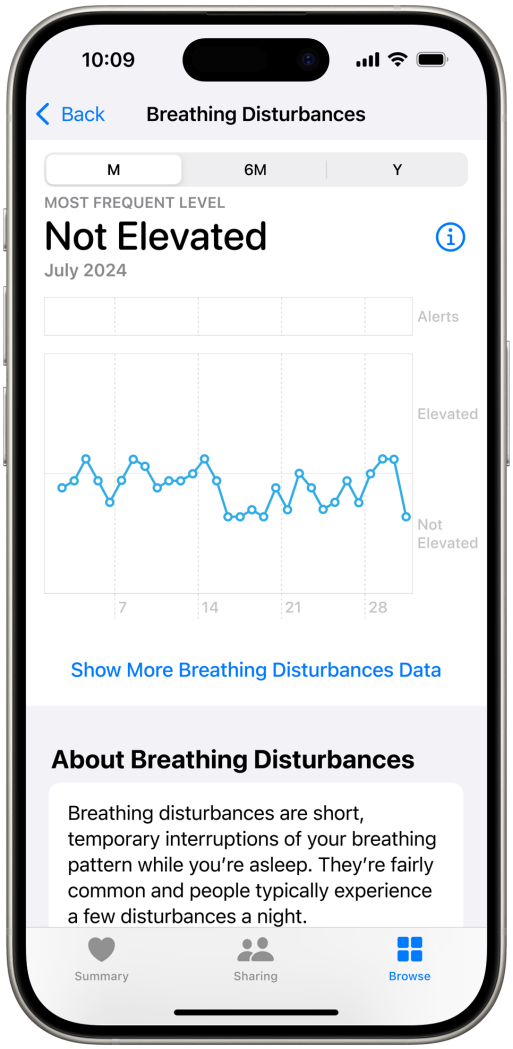


Figure 2. Breathing Disturbance Data in the Health App on iPhone

To learn more about using the Sleep feature in the Health app on Apple Watch, visit support.apple.com/en-us/HT211685.

Preclinical Design and Algorithm Testing

Apple conducted research studies to develop the Breathing Disturbances metric and the associated sleep apnea notification algorithm. Adult participants from multiple research sites provided informed consent via protocols approved by an institutional review board (IRB). To enhance performance generalizability, Apple recruited a diverse population of research participants across various demographic factors (age, biological sex, race, ethnicity, and BMI) and evaluated sleep in both at-home and in-laboratory sleeping environments. In addition, the studies included a broad range of sleep apnea severity, from normal (fewer than five apnea and hypopnea events per hour) to severe (more than 30 events per hour).

Adult research participants wore an Apple Watch while reference recordings were conducted simultaneously from in-laboratory PSG (one night) or at-home HSAT recordings (one to four nights). In each case, the reference device recordings followed a standard clinical approach. Certified PSG technologists scored the clinical recordings according to American Academy of Sleep Medicine (AASM) standards to provide reference labels of each apnea and hypopnea event. It's important to note that current AASM clinical standards allow multiple definitions for scoring sleep apnea events. Both the training and validation studies used the strictest scoring definition, which requires a 4% oxygen desaturation for hypopneas.

The design phase of algorithm development consisted of 3936 nights of at-home and in-lab PSG recordings from 2160 participants. Some participants contributed multiple nights of recordings. The performance testing phase included an additional 7220 nights from 2542 participants (see Table 1). None of the data from the testing set was used in the design phase. That is, the testing set was sequestered and unseen by the algorithm during the training phase. Participants self-reported race and ethnicity in the design and testing sets, respectively, as White (70.7% and 69.1%), Black (9.0% and 8.9%), Asian (11.4% and 13.5%), and Hispanic (7.2% and 7.6%).

The algorithm output of Breathing Disturbances is expressed as a continuous variable, which has units of events per hour. The sleep apnea notification algorithm assesses the Breathing Disturbance data every 30 days (non-rolling), starting 30 days after onboarding. When at least 10 sleep recordings (not required to be sequential) with Breathing Disturbances values occur within a given 30-day period, the notification algorithm checks if at least 50% of these values are elevated. If so, the algorithm surfaces a notification of possible sleep apnea to the user; otherwise, it remains silent. The algorithm also stays silent if there are fewer than 10 nights with a Breathing Disturbances value in a 30-day window, as the data is insufficient for analysis. Whether or not a notification is surfaced, the sleep apnea notification algorithm will remain active and attempt to analyze data after each 30-day window.

The operating point on a receiver operating characteristic (ROC) curve reflects the trade-offs incurred when choosing a threshold for a binary classifier to balance sensitivity and specificity goals. *Sensitivity*, or the true positive rate, refers to the percentage of participants with moderate to severe sleep apnea who are correctly identified by the algorithm. *Specificity* refers to the percentage of those without moderate to severe sleep apnea who wouldn't receive a notification. The operating point on the ROC curve was intentionally chosen to favor specificity, understanding that operating point positions with high specificity have lower sensitivity. This choice supported the goal of minimizing the false positive risk, which is particularly important for a feature designed to repeatedly check for signs of possible sleep apnea over time (and in consideration of the cumulative false positive rate) while simultaneously maintaining an impactful true positive rate.

In the sequestered algorithm testing data set, notification performance was 66.6% for sensitivity and 95.9% for specificity.

To understand potential confounders that could impact algorithm performance, additional regression analyses were conducted. The results suggested no significant impact of sex, BMI, race, ethnicity, periodic limb movement index (PLMI), co-sleeper status, or sleep efficiency on algorithm performance.

	Design	Testing
Participants (4702)	2160	2542
Nights (11,156)	3936	7220
Age (years)	49.1 (16.6)	48.7 (15.1)
Sex (% female)	53	50
BMI (kg/m²)	29.6 (7.3)	29.4 (7.1)
Sleep Apnea Categories		
Normal (AHI < 5)	1425	1143
Mild (AHI ≥ 5 to < 15)	423	617
Moderate (AHI ≥ 15 to < 30)	202	384
Severe (AHI ≥ 30)	110	398
PLMI (% ≥ 15)*	22.0	22.2

Data is presented as mean (SD) for continuous variables and % for categorical variables. % refers to nights (not participants), and recordings from individual participants with multiple nights were used only in either the design or testing phases (a single participant was never used in both). Periodic limb movement index (PLMI) data was available for nights with PSG reference recordings (in lab). Body mass index (BMI) is calculated as units of kilograms (kg) over meters squared (m^2).

* Based on in-lab PSG studies only.

Table 1. Participant Characteristics in Preclinical Development

Clinical Validation

Apple validated the finished algorithm separately using data collected in a prospective clinical study to support submissions to the FDA and other global medical device regulators. Each participant provided written informed consent to participate in the study, defined by an IRB-approved protocol. In this study, participants wore Apple Watch for up to 30 nights and underwent a minimum of two nights of HSAT recordings with a type 3 HSAT device, which served as the ground truth for sleep apnea status. A central scoring team of certified PSG technologists scored the HSAT data according to the AASM clinical standards, using the 4% desaturation hypopnea scoring rule. Each HSAT night was scored separately by two technologists who were blinded to each other and the algorithm. A certified sleep clinician, blinded to the algorithm, reviewed the raw data and the two technologists' scoring reports, adjudicating any disagreement between scorers using pre-specified escalation rules. Ground truth status for notification performance was defined as the higher value of two HSAT nights.

The characteristics of the participants in clinical validation are given in Table 2.

	Validation
Participants (no.)	1499
Nights (no.)	24 (7)
Age (years)	46 (14)
Sex (% female)	56.5
BMI (kg/m^2)	32 (8)
Sleep apnea categories	
Normal (AHI < 5)	559
Mild (AHI ≥ 5 to < 15)	362
Moderate (AHI ≥ 15 to < 30)	216
Severe (AHI ≥ 30)	201

Data is presented as mean (SD) for continuous variables and % for categorical variables. Body mass index (BMI) is calculated as units of kilograms (kg) over meters squared (m^2).

Table 2. Participant Characteristics in Clinical Validation

The study population included broad representation across demographic factors and sleep apnea severity, including an enriched population that exceeds the real-world prevalence of approximately 10% for moderate to severe sleep apnea, based on the strict 4% scoring rule. Therefore, to reflect the real-world sensitivity and specificity of the notification algorithm, the performance of each group was re-weighted based on approximate prevalence data in adults. The rationale for pre-specifying weights is informed by epidemiology research (Franklin and Lindberg 2015; Benjafield et al. 2019), which shows that normal AHI values are by far the most common, followed by mild-range AHI values. To compute overall specificity for the combined group of normal and mild cases, analysis included a conservative five-to-one weighting ratio. In contrast, the prevalence imbalance between moderate and severe is much less pronounced; thus, sensitivity was calculated using a simple weighting ratio of one-to-one for moderate and severe cases.

The primary endpoints for the notification performance are:

- Sensitivity of the notification for those with a ground truth indication of moderate or severe sleep apnea
- Specificity of the notification for those with a ground truth indication of normal or mild sleep apnea

The ground truth sleep apnea category for each participant was defined for the primary endpoint using the higher AHI value of two HSAT nights.

The primary endpoint hypothesis associated with the notification accuracy was pre-specified as follows:

- H_0: sensitivity < 50% or specificity < 85%
- H_1: sensitivity ≥ 50% and specificity ≥ 85%

To test the null hypothesis, H_0, estimates of the overall sensitivity and overall specificity were computed using the estimates of the individual diagnostic category performance and weights as outlined above. H_0 was rejected if the lower confidence bounds for sensitivity and specificity were at least 50% and 85%, respectively.

A total of 1499 participants were enrolled, with 1448 completing the study. The sensitivity was 66.3% (95% CI: 62.2% to 70.3%), and the specificity was 98.5% (95% CI: 98.0% to 99.0%), demonstrating that the feature meets the design objectives to confidently identify sleep apnea while minimizing false positives. Table 3 shows the performance of specificity (for normal and mild categories) and sensitivity (for moderate and severe categories). It's important to note that the specificity was 100% (95% CI: 99.7 to 100%) for the normal category, indicating that all participants with a "positive" algorithm result had at least mild sleep apnea. Also, sensitivity was higher in the severe category at 89.1% (95% CI: 83.7% to 93.2%), indicating that the large majority of severe cases were identified.

Subgroup analyses indicated that the sensitivity and specificity were comparable when analyzed by age, sex, race, ethnicity, and BMI subgroups. No serious adverse events were observed.

Algorithm Result	Value	Two-Sided 95% Confidence Interval
Sensitivity for moderate category	89/205 (43.4%)	(36.5%, 50.5%)
Sensitivity for severe category	164/184 (89.1%)	(83.7%, 93.2%)
Weighted overall sensitivity	**66.3%**	**(62.2%, 70.3%)**
Specificity for normal category	543/543 (100.0%)	(99.3%, 100.0%)
Specificity for mild category	315/346 (91.0%)	(87.5%, 93.8%)
Weighted overall specificity	**98.5%**	**(98.0%, 99.0%)**

Weighted average sensitivity = (1/2)(moderate sensitivity) + (1/2)(severe sensitivity)
Weighted average specificity = (5/6)(normal specificity) + (1/6)(mild specificity)

Table 3. Primary Endpoints: Sensitivity and Specificity

A secondary endpoint analysis was conducted on the accuracy of individual night Breathing Disturbances values, where ground truth was defined by the HSAT reference device compared with participants' aligned time wearing Apple Watch. Figure 3 shows the evaluation method for assessing performance. The green funnel-shaped zone is based on the principle that error tolerance grows as the true AHI value grows, similar to the approach used for blood sugar (Jendrike et al. 2017). For example, an error of 3 points is less meaningful when the true AHI is 50 (severe sleep apnea) than when the true AHI is 5 (borderline mild sleep apnea). A percentage-based error is also deemed unsuitable because it would place undue restrictions at a very low AHI. For example, a Breathing Disturbances value of 2, when the true AHI = 1, meaning an error of +1, would be a 100% error. The green funnel bounds are symmetric around the identity line. Of the 1305 participants who had at least one paired watch and an HSAT data value, 1193 (91.4%) had Breathing Disturbances measurements within the green performance zone.

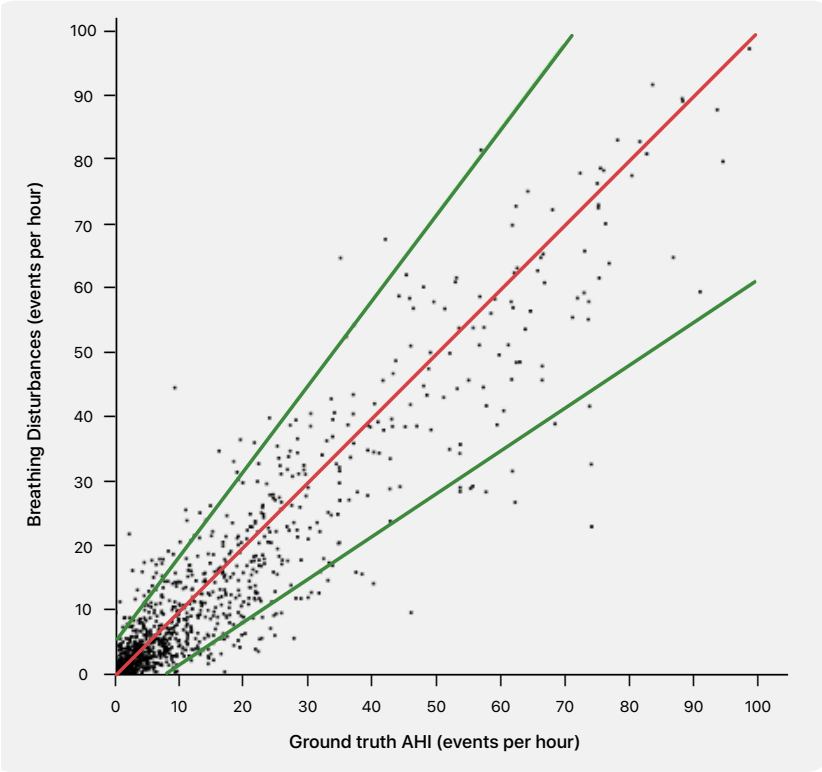


Figure 3. Funnel Region to Assess Breathing Disturbances Accuracy

Discussion

The Sleep Apnea Notification Feature is a software-based medical device that analyzes Breathing Disturbance data collected nightly by the Apple Watch accelerometer sensor. It surfaces a notification if there's sufficient information suggesting moderate to severe sleep apnea (based on multiple elevations over a 30-day period containing at least 10 days of Breathing Disturbances values). Tracking Breathing Disturbances during sleep can also help users discover patterns over time that may provide insight into the quality of their sleep. For example, various lifestyle factors, such as alcohol, exercise, and changes in medications or body weight, can influence Breathing Disturbances levels over time. The performance accuracy of this notification was validated in engineering studies and confirmed to meet performance requirements in a separate formal clinical validation study submitted to the FDA and other global medical device regulatory authorities as part of the software marketing authorization.

The sleep apnea algorithm is trained and tested on large and diverse populations, and the details provided in this report can help guide users and their care providers toward confident interpretations. This feature doesn't take into account a user's symptoms, and the Health app's user interface (UI) makes it clear that if users are experiencing symptoms, they should speak to their physician. The Health app also clarifies that the absence of a notification isn't intended to indicate the absence of sleep apnea. Most notably, the feature has the capacity to accurately track sleep apnea risk over time in ways not possible with traditional means of breathing assessments, such as PSG and HSAT, which focus on only one or a few nights.

Conclusion

Apple Watch includes a wide range of features that focus on health, fitness, safety, and staying connected. The Sleep Apnea Notification Feature sends users a notification if multiple nights of elevated Breathing Disturbances values suggest possible sleep apnea. The Health app provides an informative report on iPhone for users to discuss with their healthcare provider. Breathing Disturbance data available in the Health app also enables users to relate their lifestyle factors to their quality of sleep.

References

Epstein, Lawrence J., David Kristo, Patrick J. Strollo Jr., Norman Friedman, Atul Malhotra, Susheel P. Patil, Kannan Ramar, Robert Rogers, Richard J. Schwab, and Edward M. Weaver. 2009. "Clinical Guideline for the Evaluation, Management and Long-Term Care of Obstructive Sleep Apnea in Adults." *Journal of Clinical Sleep Medicine* 5, no. 3 (June): 263–276. pubmed.ncbi.nlm.nih.gov/19960649.

Benjafield, Adam V., Najib T. Ayas, Peter R. Eastwood, Raphael Heinzer, Mary S. M. Ip, Mary J. Morrell, Carlos M. Nunez, Sanjay R. Patel, Thomas Penzel, and Jean-Louis Pépin. 2019. "Estimation of the Global Prevalence and Burden of Obstructive Sleep Apnoea: A Literature-Based Analysis." *Lancet Respiratory Medicine* 7, no. 8 (August): 687–698. doi.org/10.1016/S2213-2600(19)30198-5.

Peppard, Paul E., Terry Young, Jodi H. Barnet, Mari Palta, Erika W. Hagen, and Khin Mae Hla. 2013. "Increased Prevalence of Sleep-Disordered Breathing in Adults." *American Journal of Epidemiology* 177, no. 9 (May): 1006–1014. doi.org/10.1093/aje/kws342.

Franklin, Karl A. and Eva Lindberg. 2015. "Obstructive Sleep Apnea Is a Common Disorder in the Population—A Review on the Epidemiology of Sleep Apnea." *Journal of Thoracic Disease* 7, no. 8 (August): 1311–1322. doi.org/10.3978/j.issn.2072-1439.2015.06.11.

Jendrike, Nina, Annette Baumstark, Ulrike Kamecke, Cornelia Haug, and Guido Freckmann. 2017. "ISO 15197: 2013 Evaluation of a Blood Glucose Monitoring System's Measurement Accuracy." *Journal of Diabetes Science and Technology* 11, no. 6 (November): 1275–1276. doi.org/10.1177/1932296817727550.