Estimating Sleep Stages from Apple Watch

September 2023

Overview

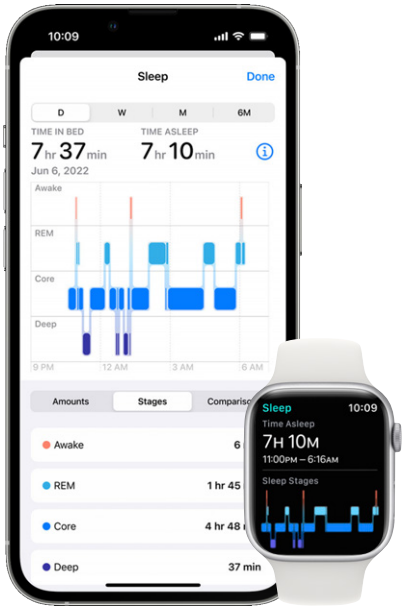
Apple Watch users can track their sleep with Apple Watch Series 3 or later using the Sleep app. The initial sleep-tracking feature, released in 2020 (with iOS 14 and watchOS 7), included a determination of two states — asleep or awake — for each 30-second window of analysis. An updated feature, released in 2022 (with iOS 16 and watchOS 9), expanded this capability to identify four states — Awake and three stages of sleep: Deep sleep, Core sleep, and REM (rapid eye movement) sleep. This paper provides a detailed understanding of the development and validation of this updated sleep stages feature.

Introduction

Sleep is a well-accepted pillar of health and well-being. Organizations such as the American Academy of Sleep Medicine (AASM),[1] the Centers for Disease Control and Prevention (CDC),[2] and the American Heart Association (AHA)[3] all highlight the importance of sleep. Tracking sleep helps users understand their patterns and the factors that may affect their sleep.

The gold standard for scoring sleep stages is polysomnography (PSG), which involves sensors on various locations on the body to capture electroencephalography (EEG), electrocardiogram (ECG), pulse oximetry, airflow, and leg movements. Sleep stages are assigned by human experts based on visual inspection of the physiological sensor measurements. Using PSG scoring rules, five stages are defined: awake, REM, and three types of non-REM sleep.[4] The non-REM stages include stage N1 (light or transitional sleep), stage N2 (the majority of non-REM sleep), and stage N3 (deep or slow-wave sleep). REM sleep is distinct from non-REM sleep in three main ways: the namesake rapid eye movements, changes in respiration pattern, and an EEG pattern that resembles the awake state.

Apple developed a sleep staging classifier algorithm that's based on accelerometer signals derived from Apple Watch. The accelerometer signals contain information about oscillations or changes that correspond to body movements, including respiration-induced motion patterns. This document describes the performance evaluation of the sleep stage classifier, which is based on algorithm training and testing (also known as design and validation) using a combination of ground truth recordings of standard PSG in the laboratory, PSG at home, and EEG at home.



The Sleep app experience requires a short onboarding process, which includes setting a sleep schedule. Once sleep tracking is enabled, and the user wears their Apple Watch to bed with Sleep Focus on, they can view their sleep stages in the Health app on iPhone or directly on Apple Watch. To the right are examples of what a user might see.

To learn more about Sleep, visit support.apple.com/en-us/HT211685.

Accelerometer-Based Sleep Staging

Technical and Feature Description
Apple Watch contains 3-axis accelerometer signals that track motion. They record not only large movements that would be visible to the naked eye but also more subtle movements, including those generated from breathing. This time series accelerometer data serves as the input to an algorithm that classifies the signal every 30 seconds (epochs) into one of four stages: Awake, REM sleep, Deep sleep, or Core sleep. This classification output aligns with standard human scoring conventions for sleep staging, as described below.

To train and test the algorithm, a combination of laboratory and at-home overnight PSG recording sessions — annotated by human experts (PSG technologists) — was used to assign a ground truth stage label for every 30-second epoch. In addition, at-home testing was performed using an EEG-based device with automated scoring followed by confirmatory human expert review. The design and validation of the algorithm is described in the next section and in Table 1. The output labels of Awake and REM refer to the corresponding clinically defined stages of the same names. The output label Deep corresponds to the clinically defined stage called N3 (for non-REM stage 3). The output label Core corresponds to the N1 and N2 stages. N1 typically accounts for only a small portion of a night of sleep. In the context of the Sleep app, Core is similar to what some sleep trackers may refer to as "light sleep." The label Core was chosen to avoid possible unintended implications of the term *light*, because the N2 stage is predominant (often making up more than 50 percent of a night's sleep), normal, and an important aspect of sleep physiology, containing sleep spindles[5] and K-complexes.[6]

Apple Sleep State	Description
REM	Rapid eye movement sleep
Deep	Stage N3 or slow-wave sleep
Core	Stages N1 and N2 sleep

Algorithm Development and Validation

Apple conducted studies to develop the sleep stages classification algorithm and to evaluate the algorithm's performance. To ensure generalizability of the algorithm, algorithm development included a diverse population of research participants — across a range of ages, sexes, races, ethnicities, and body mass indexes (BMI) — in both at-home and laboratory sleeping environments. The population included individuals without sleep disorders and individuals with sleep abnormalities, such as sleep apnea and periodic limb movements of sleep.

Adult participants provided informed consent through protocols approved by an institutional review board involving multiple sites. They underwent ground truth monitoring with either at-home or in-lab sleep recordings while simultaneously wearing an Apple Watch. PSG recordings in the lab and at home used a standard clinical montage for sleep staging, cardiorespiratory monitoring, and limb movements. Sleep staging was performed by certified PSG technologists according to AASM scoring standards.[7] At-home EEG-based recordings were performed with the Prodigy Sleep Monitor (Cerebra Health), which is autoscored and then manually overscored.

A total of 1171 nights of at-home and in-lab PSG recordings from 858 participants were used to develop the algorithm in the design phase. An additional 299 nights from 166 participants — who weren't included in the development data set — were used to assess performance and validate the algorithm (see Table 1). Participants self-reported their race and ethnicity as White (74.9 percent and 75.3 percent), Black (20.9 percent and 19.1 percent), Asian (2.0 percent and 3.7 percent), and Hispanic (5.3 percent and 4.8 percent) in the algorithm design and validation sets, respectively.

The validation set comprised subjects who were sequestered (unseen) during model training and checkpoint evaluation, and were instead utilized exclusively for performance evaluation. Some subjects contributed multiple nights of recording, and some subjects wore an Apple Watch on each wrist simultaneously. 75 percent of the sleep sessions used in design, and 78 percent of those used in validation, were from in-lab PSG recordings that spanned one to three nights; 9 percent of sleep sessions in design, and 22 percent of those in validation, were from at-home PSG recordings that spanned one to two nights. The at-home EEG recordings were used only in design, not in validation, and 16 percent of the sleep sessions in the design set were made up of at-home EEG with one to seven nights of recording.

Table 1: Subject characteristics and ground truth sleep metrics

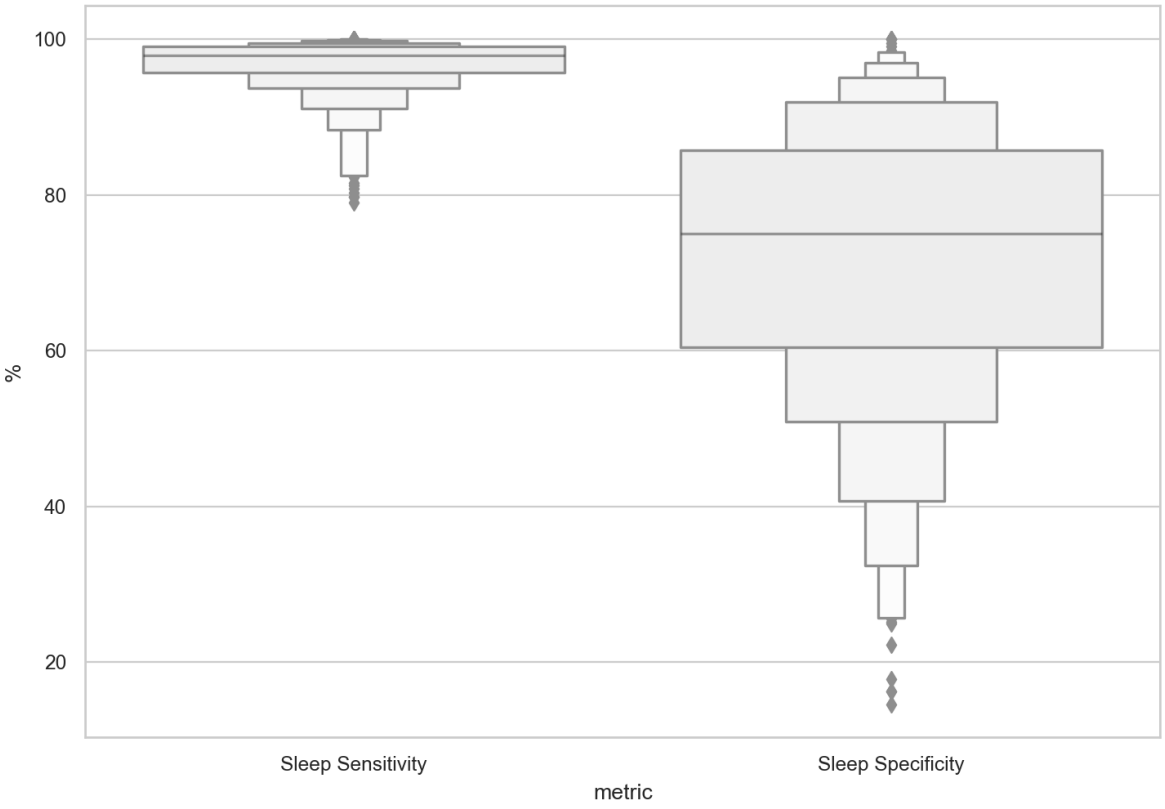
	Design	Validation
Subjects (#)	858	166
Nights (#)	1171	299
Age (years)	49 (15.9)	47 (16.5)
Sex (% female)	50	47
BMI (kg/m^2)	29.7 (7.6)	26.8 (4.7)
TST (minutes)	365 (76.2)	378 (60.7)
Efficiency (%)	80.9 (13.8)	85.4 (9.4)
Core (%)	69.5 (13.3)	65.9 (10.7)
REM (%)	18.1 (7.8)	21.4 (7.0)
Deep (%)	11.6 (9.8)	12.8 (8.7)
WASO (minutes)	31.5 (30.6)	35.5 (31.5)
Sleep apnea (category)	Mild: 16%, Moderate/severe: 15%	Mild: 11.8%, Moderate/severe: 3%
PLMI (% 15+)	19.1	13.2

Note: Data presented are means (with standard deviations in parentheses) for continuous variables and percentages for categorical variables. Percentages refer to nights (not subjects), and individual subjects with multiple nights of recording were used only in either design or validation (a single subject was never used in both phases). Efficiency was defined as total sleep time (TST) over time in bed (lights off to lights on for PSG, and recording time for home tests). Wake after sleep onset (WASO) refers to any epoch scored as wake after the first sleep epoch and before the final sleep epoch. The apnea hypopnea index (AHI) and periodic limb movement index (PLMI) were available for nights with PSG reference recordings (in lab and at home). The AHI was scored according to the 3 percent desaturation rule to qualify hypopneas, and it was used to assign the severity category: mild is >5 to <15, and moderate/severe is >=15. Body mass index (BMI) is measured in units of kilograms (kg) over meters (m) squared.

Binary sleep-wake classification performance

Many common questions in sleep health can be assessed with metrics derived from a binary classifier output that reports sleep versus wake state. In this binary setting, sleep sensitivity and specificity can be computed by collapsing Core, Deep, and REM sleeps into a single state ("Sleep"). Figure 1 shows the distribution of sensitivity and specificity values in the validation group. Each point in the distribution is a sensitivity or specificity value calculated for every session (night) of sleep recording. The median Sleep sensitivity of 97.9 percent indicates that the vast majority of true epochs of Sleep are correctly classified as Sleep, while 2.1 percent would be falsely labeled as Awake, on average. The median Sleep specificity of 75.0 percent indicates that most true Awake epochs are correctly classified as Awake, while 25.0 percent would be falsely labeled as Sleep, on average.

Figure 1: Sensitivity and specificity for the binary sleep-wake classifier

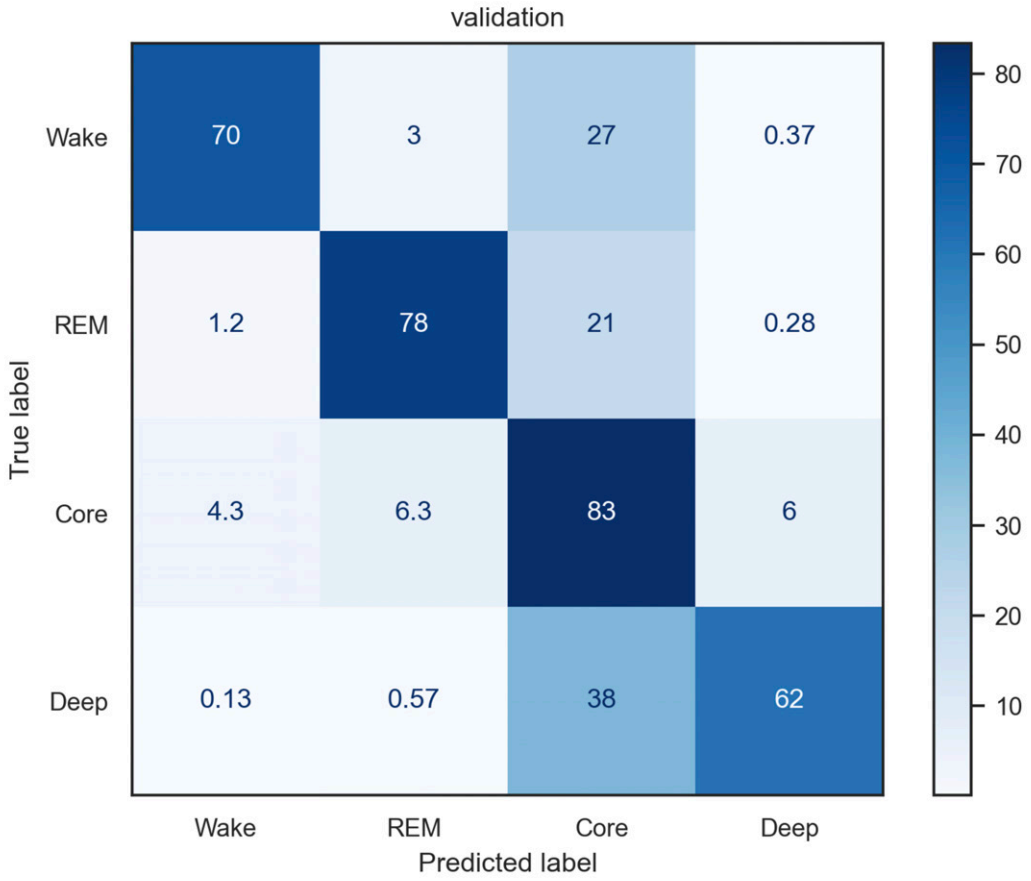


Note: This box plot visualizes the performance for binary sleep-wake classification, where sleep sensitivity (left) and specificity (right) are computed for each subject, with the distribution of performance shown across the validation cohort. The center line is the median, and the boxes above and below are the interquartile range. Subsequent boxes divide the remaining percentile range in half at each step, with outliers shown as individual points.

Four-stage classification performance

Figure 2 is a confusion matrix that shows the four-stage classification performance in the validation set across all classes, comparing ground truth labels (rows) with algorithm predictions (columns). The accuracy of classification can be seen (in the diagonal cells), as well as the proportion of each kind of classification error (off-diagonal cells). Each number is a percentage of the true epochs of that state (row) predicted by the algorithm to be in one of four states (columns). Certain classification errors, from very physiologically distinct stages, are uncommon — for example, mistaking true Deep sleep as Awake (0.13 percent), or true REM sleep as Deep sleep (0.28 percent). The most common misclassification overall was true Deep sleep being classified as Core sleep. The most common misclassification of each true stage (including Awake) was to the Core sleep stage (the lighter blue squares in the third column). The distribution of label mismatch error types is similar to the relative proportion of errors by human experts in inter-rater reliability studies.[8]

Figure 2: Four-stage classification performance



Note: This confusion matrix shows the four-stage classification performance, with true human labels in rows and Apple Watch predictions in columns. The value in each cell is the percentage of epochs scored in each possible combination for the validation cohort. The diagonal shows the percentage of true stage epochs correctly classified. The off-diagonal cells show the percentage of true sleep stage epochs classified in error. The color scale represents the percentage values in gradient.

Another way to summarize performance accuracy is through the kappa coefficient, a measure of performance accuracy that takes into account the role of chance. The average four-stage kappa value in the validation set was 0.63 (SD 0.13). 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 (coefficients of –0.037 and –0.036, respectively) impacted kappa performance, which isn't unexpected because factors that directly influence sleep architecture — such as frequent stage transitions — are predicted to be more challenging for classification. Male sex (–0.014) and age (–0.014) had proportionately lower impact on kappa, with small but statistically evident coefficients. Different Apple Watch models were assessed, and they were found to have no impact on kappa values.

Finally, a cohort of adult patients undergoing clinical PSG as part of routine care — 236 subjects — were independently assessed for performance, with 390 recorded watch sessions (some subjects wore a watch on both wrists). Although Apple Watch sleep stages aren't intended for clinical use, expanding the performance assessment to challenging populations provides greater context for how well performance generalizes in diverse consumer populations. In this separate validation set, the kappa on clinical PSG recordings was 0.55 (0.65 to 0.57), which isn't unexpected given the added complexity of clinical populations.[9] From a binary classification perspective, the clinical population validation set showed a median sleep sensitivity of 97.5 percent (0.4 percent lower than the nonclinical population) and a specificity of 72.5 percent (2.5 percent lower than the nonclinical population).

Discussion

Tracking sleep over time can help users discover patterns and increase their understanding of, and respect for, this central aspect of their lives. Sleep is also quite personal; habitual duration, timing, and perceived quality may differ among people, and they can even change over time for a given person. And one's subjective perception can be more complex than expected. For example, brief awakenings are common in normal sleep, but they may go unrecognized by the sleeper. Lifestyle factors — ranging from alcohol and caffeine consumption to exercise and stress — and schedule inconsistency may influence the timing, duration, and stages of sleep. Understanding the accuracy and limitations of sleep-tracking devices provides an important basis for adding objective metrics to this quintessentially personal experience of sleep.

Research has long suggested that sleep in general, and the REM and Deep stages specifically, are involved in various aspects of cognition,[10] among other bodily rejuvenation functions. But many questions remain unanswered when it comes to translating these findings into a health context,[11] including concepts of sleep quality.[12] Several review articles on consumer sleep trackers have been published in recent years,[13] including from the AASM,[14] covering a range of topics such as algorithm performance. Reliable sleep tracking may help bridge the gaps between research studies and real-world experiences. The Apple Watch sleep-tracking performance is trained and tested on large and diverse populations appropriate for performance validation, whether considering overall sleep-wake patterns or the stages that compose sleep. The accuracy is suitable for a variety of use cases, and the details provided here can help guide users and researchers toward confident interpretations. This includes, most notably, the capacity to accurately track sleep over time in ways not possible with traditional means of EEG-based assessments for one or a few nights, and with accuracy not possible with actigraphy.

For example, the traditional method of multinight tracking uses wrist actigraphy,[15] which derives activity counts (gross movement) from accelerometer signals. It has a relatively high sensitivity for sleep (generally 90–97 percent), while the specificity is generally much lower — often 50 percent or less.[16] The imbalance of high sleep sensitivity with relatively low specificity indicates that periods of true sleep are often correctly labeled as such, while periods of true wake are often misclassified as sleep. By comparison, Apple Watch has a very high sleep sensitivity (97.9 percent in the validation set) and a high specificity of 75.0 percent. Unlike traditional actigraphy algorithms that use binned activity counts, the Apple Watch algorithm uses the high-frequency sampling of accelerometer data from the wrist, which carries more granular information about body movements, including respiration.[17] This may contribute to our classifier performance and capacity to distinguish stages, in comparison with traditional actigraphy.

Conclusion

Apple Watch includes a wide range of features that focus on health, fitness, safety, and connectivity. Sleep stages on Apple Watch allow users to track their sleep and make their own connections with lifestyle and environmental factors that can influence sleep. The data is also available in the Health app on iPhone for additional insights on values and trends. Features that track activity, heart rate, cardio fitness, sleep stages, and SpO2 make Apple Watch a powerful device for wellness.

1. Nathaniel F. Watson et al., "Recommended amount of sleep for a healthy adult: a joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society," *Journal of Clinical Sleep Medicine* 11, no. 6 (June 2015): 591–592, aasm.org/resources/pdf/pressroom/adult-sleep-duration-consensus.pdf.
2. "Sleep: About Our Program," Sleep and Sleep Disorders, Centers for Disease Control and Prevention, last reviewed December 14, 2022, www.cdc.gov/sleep/about_us.html.
3. "Life's Essential 8™ - How to Get Healthy Sleep Fact Sheet," Healthy Living: Healthy Lifestyle, American Heart Association, May 2022, www.heart.org/en/healthy-living/healthy-lifestyle/lifes-essential-8/how-to-get-healthy-sleep-fact-sheet.
4. "The AASM Manual for the Scoring of Sleep and Associated Events," AASM Scoring Manual, American Academy of Sleep Medicine, 2023, aasm.org/clinical-resources/scoring-manual.
5. Laura M. J. Fernandez and Anita Lüthi, "Sleep Spindles: Mechanisms and Functions," *Physiological Reviews* 100, no. 2 (April 2020): 805–868, pubmed.ncbi.nlm.nih.gov/31804897.
6. Péter Halász, "The K-complex as a special reactive sleep slow wave - A theoretical update," *Sleep Medicine Reviews* 29, (October 2016): 34–40, pubmed.ncbi.nlm.nih.gov/26606317.
7. "The AASM Manual for the Scoring of Sleep and Associated Events," AASM Scoring Manual, American Academy of Sleep Medicine, 2023, aasm.org/clinical-resources/scoring-manual.
8. Heidi Danker-Hopfe et al., "Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders," *Journal of Sleep Research* 13, no. 1 (March 2004): 63–69, pubmed.ncbi.nlm.nih.gov/14996037.
9. Heidi Danker-Hopfe et al., "Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard," *Journal of Sleep Research* 18, no. 1 (March 2009): 74–84, pubmed.ncbi.nlm.nih.gov/19250176.
10. Matthew P. Walker, "The Role of Sleep in Cognition and Emotion," *Annals of the New York Academy of Sciences* 1156, no. 1 (March 2009): 168–197, pubmed.ncbi.nlm.nih.gov/19338508; Susanne Diekelmann, Ines Wilhelm, and Jan Born, "The whats and whens of sleep-dependent memory consolidation," *Sleep Medicine Reviews* 13, no. 5 (October 2009): 309–321, pubmed.ncbi.nlm.nih.gov/19251443.
11. James Krueger et al., "Sleep function: Toward elucidating an enigma," *Sleep Medicine Reviews* 28 (August 2016): 46–54, www.ncbi.nlm.nih.gov/pmc/articles/PMC4769986.
12. Maurice Ohayon et al., "National Sleep Foundation's sleep quality recommendations: first report," *Sleep Health* 3, no. 1 (February 2017): 6–19, pubmed.ncbi.nlm.nih.gov/28346153.
13. Massimiliano de Zambotti et al., "Sensors Capabilities, Performance, and Use of Consumer Sleep Technology," *Sleep Medicine Clinics* 15, no. 1 (March 2020): 1–30, ncbi.nlm.nih.gov/pmc/articles/PMC7482551; Massimiliano de Zambotti et al., "Rigorous performance evaluation (previously, 'validation') for informed use of new technologies for sleep health measurement, *Sleep Health* 8, no. 3 (June 2022): 263–269, pubmed.ncbi.nlm.nih.gov/35513978; Christopher M. Depner et al., "Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions," *Sleep* 43, no. 2 (February 2020): zsz254, www.ncbi.nlm.nih.gov/pmc/articles/PMC7368340; and Ping-Ru T. Ko et al., "Consumer Sleep Technologies: A Review of the Landscape," *Journal of Clinical Sleep Medicine* 11, no. 12 (December 2015): 1455–1461, www.ncbi.nlm.nih.gov/pmc/articles/PMC4661339.
14. Seema Khosla et al., American Academy of Sleep Medicine Board of Directors, "Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement," *Journal of Clinical Sleep Medicine* 14, no. 5 (May 2018): 877–880, www.ncbi.nlm.nih.gov/pmc/articles/PMC5940440; and Sharon Schutte-Rodin et al., "Evaluating consumer and clinical sleep technologies: an American Academy of Sleep Medicine update," *Journal of Clinical Sleep Medicine* 17, no. 11 (November 2021): 2275–2282, pubmed.ncbi.nlm.nih.gov/34314344.
15. Jennifer L. Martin and Alex D. Hakim, "Wrist Actigraphy," *Chest* 139, no. 6 (June 2011): 1514–1527, www.ncbi.nlm.nih.gov/pmc/articles/PMC3109647; and Michael T. Smith et al., "Use of Actigraphy for the Evaluation of Sleep Disorders and Circadian Rhythm Sleep-Wake Disorders: An American Academy of Sleep Medicine Clinical Practice Guideline," *Journal of Clinical Sleep Medicine* 14, no. 7 (July 2018): 1231–1237, www.ncbi.nlm.nih.gov/pmc/articles/PMC6040807.
16. Avi Sadeh, "The role and validity of actigraphy in sleep medicine: An update," *Sleep Medicine Reviews* 15, no. 4 (August 2011: 259–267, pubmed.ncbi.nlm.nih.gov/21237680; and Michael T. Smith et al., "Use of Actigraphy for the Evaluation of Sleep Disorders and Circadian Rhythm Sleep-Wake Disorders: An American Academy of Sleep Medicine Systematic Review, Meta-Analysis, and GRADE Assessment," *Journal of Clinical Sleep Medicine* 14, no. 7 (July 2018): 1209–1230, www.ncbi.nlm.nih.gov/pmc/articles/PMC6040804.
17. Julian Leube et al., "Reconstruction of the respiratory signal through ECG and wrist accelerometer data," *Scientific Reports* 10 (September 2020): 14530, pubmed.ncbi.nlm.nih.gov/32884062; and Marian Haescher, Denys J. C. Matthies, John Trimpop, and Bodo Urban, "A study on measuring heart- and respiration-rate via wrist-worn accelerometer-based seismocardiography (SCG) in comparison to commonly applied technologies" (in proceedings of the iWOAR '15 - 2nd international Workshop on Sensor-based Activity Recognition and Interaction, Association for Computing Machinery, New York, NY, June 2015), article 2: 1–6, dl.acm.org/doi/10.1145/2790044.2790054.

© 2023 Apple Inc. All rights reserved. Apple, the Apple logo, Apple Watch, iPhone, and watchOS are trademarks of Apple Inc., registered in the U.S. and other countries. IOS is a trademark or registered trademark of Cisco in the U.S. and other countries and is used under license. Other product and company names mentioned herein may be trademarks of their respective companies.