Sleep-wake Detection With a Contactless, Bedside Radar Sleep Sensing System

Michael Dixon, Logan D. Schneider, Jeffrey Yu, Jonathan Hsu, Anupam Pathak, D. Shin, Reena S. Lee, Mark Malhotra, Ken Mixter, Michael V. McConnell, James A. Taylor, Shwetak N. Patel*

*For correspondence:
Google LLC
601 N 34th St
Seattle, WA 98103
shwetakpatel@google.com

March 2021

Contents

Overview

Sleep constitutes nearly ⅓ of the human lifespan, yet most individuals are unaware of precisely how much or how well they're sleeping. With low-energy radar technology, integrated into the new second-generation Nest Hub device, users can access a contactless, bedside sleep-sensing system. Radar-based detection of sub-centimeter body movements enables the passive monitoring of sleep patterns with relative ease (e.g., no need to remember to charge the device or turn it on) and without the need for cameras, microphones, or physical contact with the user. Moreover, privacy-preserving, on-device processing of the raw sensor data is employed, so only the results of the algorithm (e.g., awake or asleep) are securely uploaded and provided to the user.

This paper provides a detailed understanding of the capabilities of the second-generation Nest Hub's Sleep Sensing feature, including algorithm development and validation. In brief, the deep learning algorithm, when compared to gold-standard clinical sleep assessment, achieved overall epoch-by-epoch sleep-wake accuracy of 87% in healthy sleepers: correctly detecting 96% of sleep epochs and 55% of wake epochs. This accuracy is comparable to published results for other clinical- and consumer-grade devices.

Introduction

Inadequate sleep has been linked to nearly half of the conditions that are the top causes of mortality in the United States.[1,2] However, due to a multitude of factors - from lack of awareness of what constitutes healthy sleep to limited access to care - many individuals end up deprioritizing sleep, resulting in accumulation of a chronic sleep debt and progressive impairment in daily function and wellbeing. This has left nearly two thirds of adults in developed nations not attaining the recommended amount of nightly sleep.[2]

Recent advances in technology have enabled sleep to be monitored on a more continuous basis through a variety of different consumer-grade sensors. Given these technologies often lack accuracy on par with the gold-standard polysomnogram (PSG), a recent American Academy of Sleep Medicine Position Statement noted that they are not currently deemed appropriate for the diagnosis or treatment of sleep disorders.[3] Nonetheless, sleep-tracking wearables and "nearables" (contactless sensors) provide individuals the opportunity to understand their sleep patterns,[4] with the potential to make adjustments to their behaviors that can promote better sleep and overall wellness.

Methods

Signal analysis and algorithm development

To automatically track sleep at a short distance from the user, the second-generation Nest Hub uses Soli (https://atap.google.com/soli/), a 60 GHz frequency-modulated continuous wave (FMCW) radar chip developed for use in consumer electronic devices.[5] It operates by emitting an ultra-low-power radio wave and measuring its reflection from the region of interest. The frequency spectrum of the reflected signal contains an aggregate representation of the distance and velocity of objects within the scene. This signal can be processed to isolate a specified range of distances corresponding to the user's sleeping area and detect motion within this region.

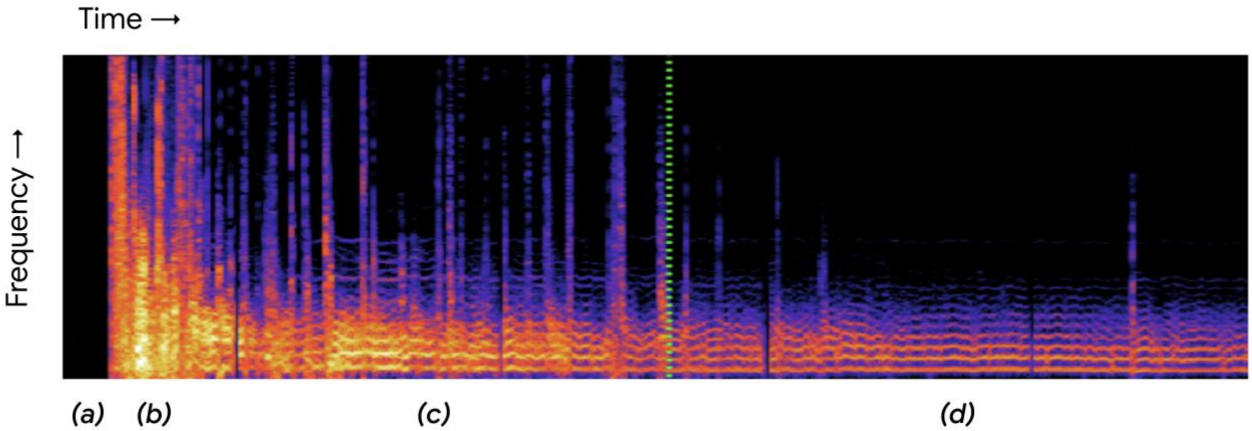

Figure 1. A Soli spectrogram demonstrating the ability to detect a wide range of motions. Following the target user entering the scene, the spectrogram helps to differentiate (a) an empty room (no variation in the reflected signal demonstrated by the black space) from (b) large pose changes, (c) brief, isolated limb movements, and (d) sub-centimeter chest and torso displacements from human breathing while at rest.

As shown in Figure 1, once a user enters the target space causing variations in the reflected radar signal, the Soli chip is sensitive enough to detect and characterize a wide range of motions, deriving full-body actigraphy from large pose changes to smaller limb movements, and even exploring cardiopulmonary physiology from sub-centimeter chest and torso displacements during quiet respiration. These clearly recognizable patterns help determine whether or not a person is present in the specified area and, if so, whether the person is asleep or awake.

To perform this task, a convolutional neural network (CNN) was developed to distinguish between three possible states: absent, awake, and asleep. This classifier was trained from a large, supervised dataset comprising raw radar data, participant-provided sleep diaries, reference sensor recordings, and external annotations collected from thousands of volunteers. At inference time, raw radar data is processed to produce a continuous series of 3D tensors representing the amount of activity across a range of distances and frequencies and over a given window of time. These features are classified by the CNN to produce a time series of state probability estimates, which are further processed to determine the most likely state (absent, awake, or asleep).

Algorithm validation

To validate the system’s performance, a reference dataset was collected, comprising raw sensor data along with corresponding ground-truth, which allows us to quantitatively measure the accuracy of the current algorithm and subsequently compare its performance to any future versions. To prevent overfitting to this benchmark, this study data was never used during model training or algorithm development and was not used to inform any design decisions; validation was only performed to obtain final performance metrics after algorithm designs were completed.

The primary ground truth was in-lab polysomnography (PSG), which was scored by two board-certified PSG technologists. In order to ensure an accurate ground truth, only epochs classified by both technologists as sleep (or wake) were used in the epoch-by-epoch analysis. Epoch-by-epoch analyses in the sleep studies were done using 30-second epochs, comparing the algorithm’s prediction of sleep and awake/absent to the set of epochs that were similarly scored by both scorers (i.e., concordant epochs). Similarly, because inter-scorer reliability for the two scorers was good (in excess of the traditional standards for clinical benchmarks[6]), comparison was made to the average of summary sleep metrics generated by each of the scorers, in order to eliminate bias in the ground truth that might be introduced by either scorer’s differing application of the rules.

Validation study population

The validation study was conducted in order to assess performance of the Sleep Sensing algorithm in a convenience sample of volunteers recruited from clinical sleep disorder populations of SleepMed (a national healthcare and clinical research organization, and subsidiary of BioSerenity) and the surrounding community of South Carolina. The validation study was approved by the Advarra institutional review board, and all participants provided written informed consent prior to participation.

Participants were recruited through multimedia advertising and direct contact of previously studied individuals who had authorized future recruitment. Participants were recruited to fulfill quotas that ensured uniform representation of certain demographic groups - ages 18-80 and male/female sex - that are likely to use the device (Table 1). Healthy sleeper participants were recruited through either evidence (on PSG) of no significant sleep disorders, or a combination of screening measures (e.g., Insomnia Severity Index, Epworth sleepiness scale) and self-report that suggested low pre-test probability for sleep disorders at the time of intake. Finally, all subjects were asked to abstain from the consumption of alcohol, caffeine, or nicotine containing products prior to the sleep study. In order to ensure natural sleep physiology, the following exclusion criteria were applied:

1. Regular shift work or any night shift work in the past 4 weeks
2. Travel across more than two time zones in the past 3 weeks
3. Serious health conditions such as severe/decompensated cardiovascular disease, history of myocardial infarction or stroke, respiratory conditions (e.g., emphysema, chronic bronchitis), or any other conditions deemed by the principal investigator to be serious, severe, or decompensated
4. Current pregnancy

5. Current or recent history (within 3 months) of major psychiatric disorders (e.g., severe or uncontrolled requiring initiation of treatment with a medication or hospitalization), history of schizophrenia or bipolar I disorder, or drug dependency
6. Any use of the following prescribed medications within the past 3 months (as determined by self-report or study questionnaires) of:
    a. antipsychotics (e.g., haloperidol, prochlorperazine, aripiprazole, etc.)
    b. mood stabilizers (e.g., lithium, carbamazepine, lamotrigine, valproate, asenapine)
    c. sleep medications (e.g., TCAs, benzodiazepines, “Z”-drugs, barbiturates, etc.)
    d. anticholinergic medications (Detrol, Elavil, etc.)
    e. opiate/opioid analgesics (Demerol, Percocet, etc.)
    f. wake-promoting therapeutics (Adderall, Ritalin, etc.)
7. Symptoms of active illness (e.g., fever) on the night of the study visit

Polysomnography

The participants in the validation cohort underwent a single night of standard, level I PSG in a clinical research facility operated by BioSerenity. The PSG included all American Academy of Sleep Medicine (AASM) required signals - electroencephalography, electrooculography, chin and tibialis anterior electromyography, nasal pressure transducer, naso-oral thermistor, snore microphone, electrocardiography, thoracic and abdominal respiratory inductance plethysmography, digital photoplethysmography/oximetry - sampled at the recommended frequencies and stored digitally.[7] In a sleep-technologist monitored setting, participants were allowed to sleep in a temperature-controlled environment with a bedtime of their choosing (usually starting around 21:00-23:00) and were awoken approximately 8 hours later at a preselected time.

Using the AASM scoring manual,[7] staging and scoring of each PSG was performed by two board-certified sleep PSG technologists, independently. Standard summary metrics were extracted from the scored files including sleep duration (total sleep time [TST]), the time from lights off to lights on (time in bed [TIB]), time of sleep onset (SOn; the first epoch of sleep following lights off), time from lights off to sleep onset (sleep onset latency [SOL]), sleep offset (SOff; the last epoch of sleep), time spent awake in between sleep onset and offset (wake after sleep onset [WASO]), and percentage of time in bed spent sleeping in between lights off and lights on (sleep efficiency [SE]). Additionally, standard indices of sleep phenomena were calculated as events per hour of sleep: periodic limb movement index (PLMI), apnea-hypopnea index using the AASM-recommended hypopnea criterion of associated 3% desaturation or electrographic arousal (AHI), and apnea-hypopnea index using the alternate hypopnea criterion of associated 4% desaturation that is commonly used for Medicare patients (CMS-AHI).

Statistical analysis

Descriptive statistics were assessed for normality using the Shapiro-Wilk test. Following confirmation of normality, continuous data were presented as mean±SD or, in the case of comparisons of sleep summary metrics mean (95% CI). Non-normal variables were represented by median and interquartile range. Count data were summarized as N(%).

Epoch-by-epoch comparisons resulted in the generation of a number of standard statistics for each individual/night of sleep. The performance metrics of interest were sensitivity, specificity, positive predictive value, negative predictive value, and accuracy. For purposes of calculating these metrics a true positive (TP) indicates an epoch classified as sleep by the algorithm and as any stage of sleep (N1, N2, N3, REM) by both technicians. Conversely, a false negative (FN) occurred when the algorithm classified such epochs as wake. A true negative (TN) was counted when the algorithm classified an epoch as awake or absent and both technicians scored wake (W) or out of bed; whereas, false positives (FP) indicated an instance where the algorithm incorrectly identified sleep when the technicians both scored wake/out-of-bed.

For normally distributed variables means were compared with two-tailed *t*-tests. Significance was set at an α threshold of 0.05, without correction for multiple comparisons, as this was a pilot validation study. Analyses were performed using version 4.0.3 of the R statistical programming language.[8] Bland-Altman plots were generated using the BlandAltmanLeh package.

Results

Cohort demographics

Of the volunteers who were screened for study inclusion, 7 were excluded due to not meeting the inclusion/exclusion criteria (1 for sleep apnea, 1 for clonazepam usage, 1 for prescription stimulant usage, 1 for ISI >7, 1 for unwillingness to abstain from alcohol, 1 for prior stroke, and 1 for inability to complete consent), and 4 were withdrawn after consenting (3 no shows to study visit that were lost to follow up; 1 withdrew due to concerns over safety related to the COVID-19 pandemic). There were 33 nights of data available for analysis after removal of nights that lacked analyzable signals due to improper set-up (2 with device miscalibration, 2 with no recorded data from the Nest Hub). The PSG validation cohort was generally representative of the regional U.S. population from which it was recruited, with intentional over-representation of specific demographics (e.g., age) (Table 1). This cohort was older 41.6±11.8 and comprised of more females (61%) than the U.S. population - median age 38.5 years and 51% female[9] - and global population - average median age 29.6[10] and 50% female[11]. The cohort was generally overweight, congruent with the U.S. population.[12] The cohort was 64% White and 36% Black or African American.

The population did not have a significant degree of sleep-disordered breathing - CMS AHI 1.81±1.16; AASM AHI 6.53±3.97 - or periodic limb movements during sleep - periodic limb movement index (PLMI) of 3.96±9.67.[7] The sleep duration was generally in excess of 6 hours, constituting 81.8%±10.4% of the nearly 8 hour sleep opportunity afforded. Based on validated cut points for the Insomnia Severity Index (ISI >7)[13] and the SleepMed Insomnia Index (SMII >20),[14] as a whole, these individuals did not have clinically significant insomnia symptoms.

	Mean±SD or N(%) (N=33)
Demographic variables	
Age (yrs)	41.6±11.8
Sex (F)	20 (61%)
BMI (kg/m²)	27.05±4.24
Race/Ethnicity	
American Indian or Alaska Native	0 (0%)
Asian	0 (0%)
Black or African American	12 (36%)
Native Hawaiian or Other Pacific Islander	0 (0%)
White	21 (64%)
More than one race	0 (0%)
Unknown or not reported	0 (0%)
Clinical variables	
CMS AHI (events/hr)	1.81±1.16
AHI (events/hr)	6.53±3.97
NadirSpO2 (%)	86.3%±7.5%
TST (min)	386.99±50.33
TIB (min)	473.21±21.28
SE (%)	81.8%±10.4%
PLMI (events/hr)	3.96±9.67
ISI	3.3±2.39
SMII	5.12±3.03

Table 1. Demographics of the validation cohort. Abbreviations: AHI - AASM-recommended apnea-hypopnea index; BMI - body mass index; CMS AHI - AASM alternate AHI, commonly used by Medicare; ISI - Insomnia Severity Index; PLMI - periodic limb movement index; SE - sleep efficiency; SMII - SleepMed Insomnia Index; TIB - time in bed; TST - total sleep time

Epoch-by-epoch performance

There was no statistically significant difference in the sleep-wake detection performance between sleep technicians (κ=0.93±0.2 for 5-stage sleep scoring). A median of 4.4% [IQR: 3.2%-6.6%] of epochs were removed because of disagreement between scorers. Performance metrics were aggregated across the whole cohort, after being calculated on a per-individual basis (Table 2). The overall accuracy of the algorithm for sleep-wake detection was 0.87±0.06, with correct detection of sleep epochs (sensitivity or recall) of 0.96±0.06 and wake epochs (specificity) of 0.55±0.20. Positive and negative predictive values exceeded 85%.

	Algorithm performance
Sensitivity	0.96±0.06
Specificity	0.55±0.20
Accuracy	0.87±0.06
PPV	0.88±0.07
NPV	0.86±0.17

Table 2. Epoch-by-epoch performance of algorithm vs concordant epochs of 2 expert scorers. Abbreviations: NPV - negative predictive value; PPV - positive predictive value

In comparison to previously published performance of other sleep-tracking devices, the Nest Hub Sleep Sensing algorithm demonstrated sleep-wake detection accuracies on par with or, in some cases, better than existing clinical and consumer sleep-tracking devices (Figure 2 & Supplementary Table 1).

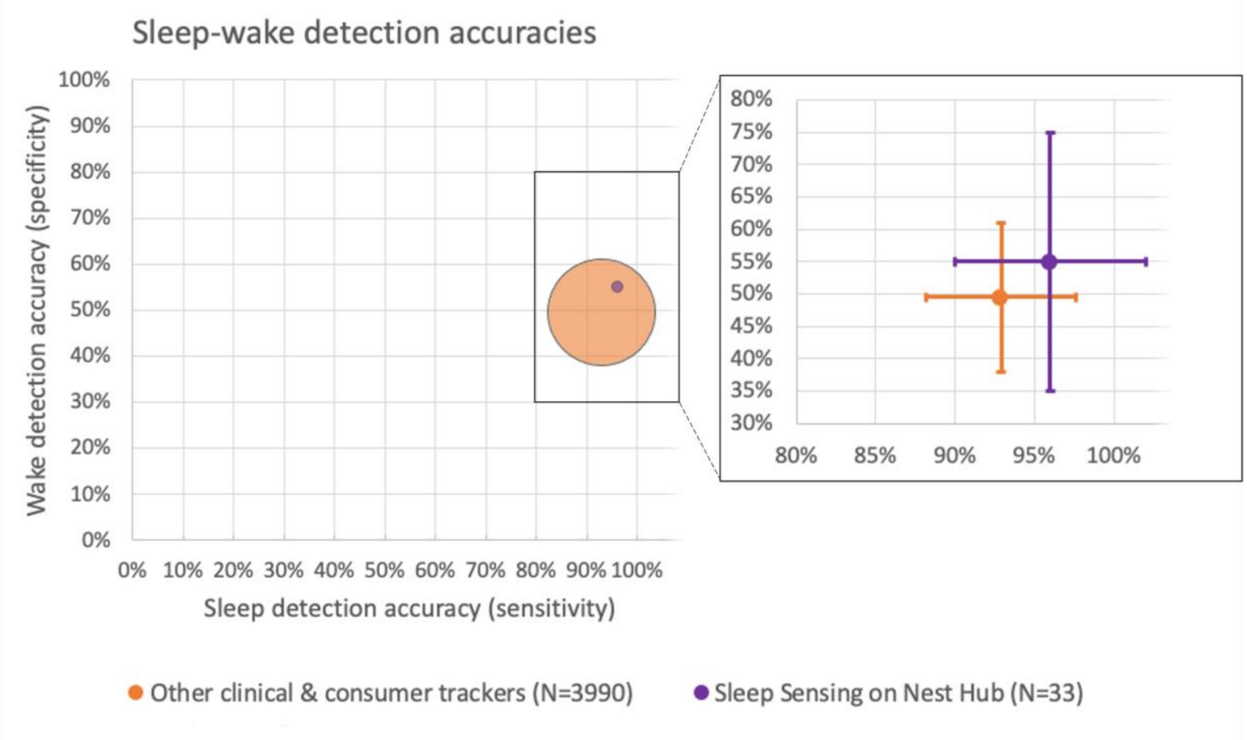

Figure 2. Comparison of the Sleep Sensing algorithm to aggregated performance of various sleep tracking technologies. Aggregate performance (in orange) from previously published accuracies for detection of sleep (sensitivity) and wake (specificity) of a variety of sleep trackers against polysomnography in a variety of different studies, accounting for 3,990 nights in total. The performance of Sleep Sensing on Nest Hub (in purple) in a population of healthy sleepers who underwent polysomnography along with the second-generation Nest Hub is added to the figure for rough comparison. The size of the circles is a reflection of the number of nights. The zoomed-in plot illustrates the means±standard deviations for the performance metrics. Perfect performance would be in the top-right corner of the figure (i.e., 100% accuracy for both sleep and wake detection).

Summary sleep metric performance

As shown in Table 3, the algorithm-determined sleep metrics were not statistically different from expert-scored PSG in the estimation of sleep onset, sleep offset, sleep onset latency, total sleep time, or sleep efficiency (*p*>0.1 for all). Comparatively, the algorithm tended to underscore wake after sleep onset by about 25 minutes (95% CI: -34.85 to -15.28; *t*-test, *p*-value: <0.001). Bland-Altman plots (Figure 3) were generated to explore patterns in the bias of algorithm performance. Notable were a 12min delay for sleep offset (11:39; 95% CI: 4:49 to 18:28); and a 35min overestimation of total sleep time, primarily related to the aforementioned underestimation in wake after sleep onset, which also resulted in an overestimation of sleep efficiency by 7%).

	Mean difference (95% CI)	**t-test, p-value**
SOn (min:sec)	1:14 (-5:48,8:15)	0.62
SOff (min:sec)	11:39 (4:49,18:28)	0.67
TST (min)	35.48 (22.13,48.84)	0.15
SOL (min)	1.23 (-5.8,8.25)	0.42
WASO (min)	-25.07 (-34.85,-15.28)	<0.01
SE (%)	7.23% (4.4%,10.06%)	0.16

Table 3. Summary metric performance of algorithm vs average of 2 expert scorers. Difference between algorithm and average of 2 expert scorers, expressed as mean (95% CI) with accompanying p-values of a two-tailed, paired t-test assessing the difference in means. Abbreviations: SE - sleep efficiency; SOff - sleep offset; SOL - sleep onset latency; SOn - sleep onset; TST - total sleep time; WASO - wake after sleep onset

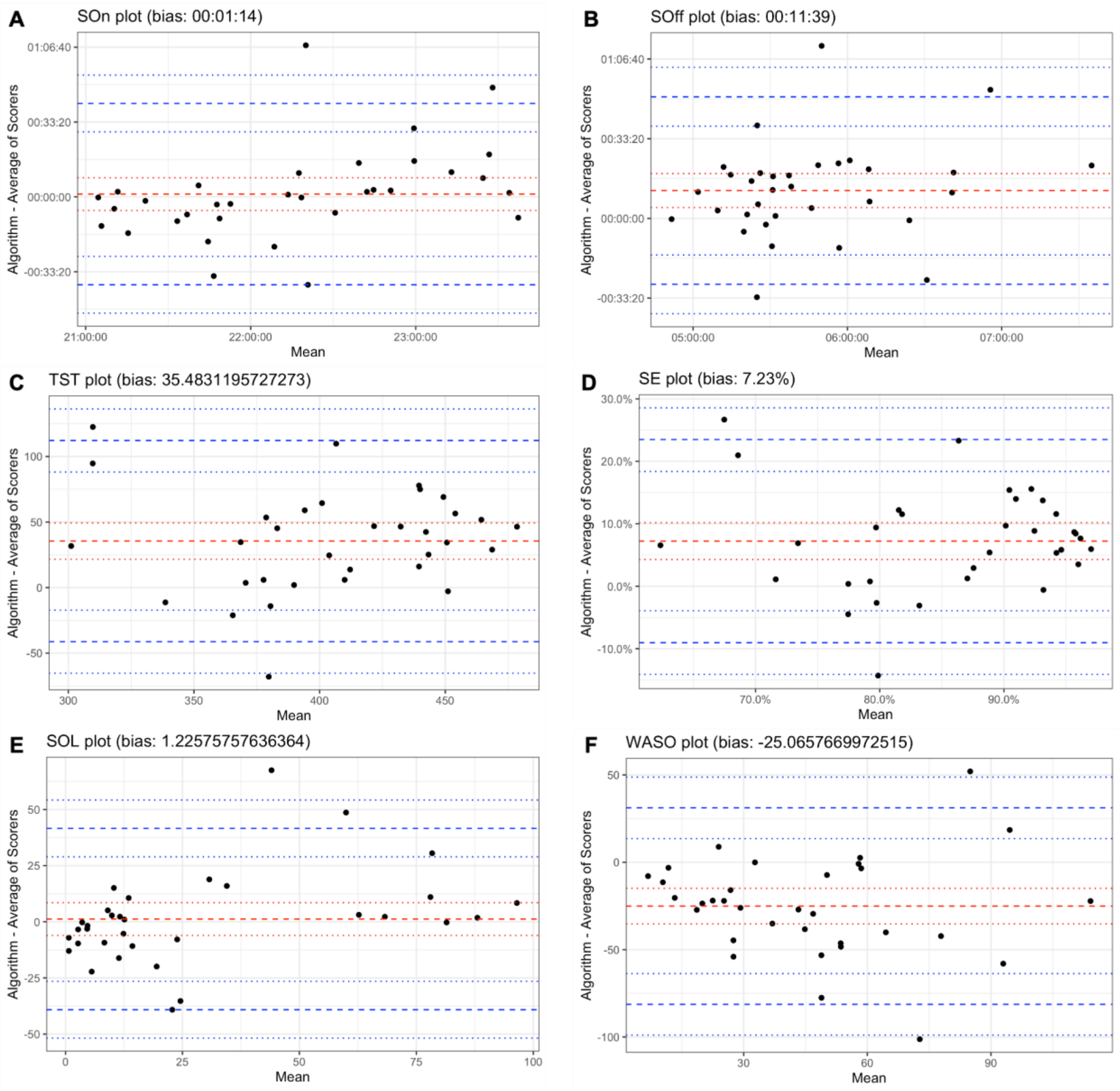

Figure 3. Bland-Altman plots of various sleep summary metrics. Comparisons between the algorithm and the average of 2 expert scorers, with the abscissa represents the mean of the algorithm and scorers and the ordinate represents the difference between the algorithm and scorers (where a positive indicates a high/delayed estimate and a negative indicates a low/early estimate). Dashed and dotted lines plotted for the mean difference and 95% CI (red) along with 2 standard deviations above/below the mean and 95% CI (blue). Abbreviations: SE - sleep efficiency; SOff - sleep offset; SOL - sleep onset latency; SOn - sleep onset; TST - total sleep time; WASO - wake after sleep onset

Conclusions

We report here on the performance of sleep-wake detection by a radar-based, bedside device and algorithm, in a cohort of 33 healthy sleepers, using clinical PSG for comparison. The overall epoch-by-epoch sleep-wake accuracy was 87%, correctly detecting 96% of sleep epochs and 55% of wake epochs. The findings for this deep learning algorithm are comparable to the reported sensitivity range (65-99%) and specificity range (10-82%) of clinical grade actigraphy, as well as many sleep-tracking devices currently used by consumers (Figure 2 & Supplementary Table 1). Along these lines, expected overestimation of sleep duration and underestimation of time spent awake occurred; however, except for underestimation of the amount of wakefulness interrupting the sleep period, analyses revealed that the algorithm's determination of major summary metrics were, on average, not statistically different than the average scoring of 2 expert sleep technicians.

The second-generation Nest Hub with Sleep Sensing allows users to monitor their sleep patterns with performance similar to that of other sleep-tracking devices.

Acknowledgements

This work involved collaborative efforts from a multidisciplinary team of software engineers, researchers, clinicians, clinical operations experts, and cross-functional contributors. In addition to the authors, key contributors to this research from Google Health include Arno Charton, Jian Cui, Jake Garrison, Navreet Gill, Sinan Hersek, Ajay Kannan, Mukil Kesavan, Linda Lei, Shao-po Ma, Kunal Okhandiar, Jo Schaeffer, Siddhant Swaroop, Lourella Palao, Allen Jiang, Bhavana Koka, Andi Janti, Neil Smith, Xiaoxia Zheng, Paolo Bonaccorsi, Sergio Balcorta, Zainab Ali, and the extended team. Also, this work would not have been possible without the dedication of our team of expert research annotators: Yanhong Li, Bhavna Daryani, Erik Schenck, Peninah Kaniu, Bryan Young, and Henry Co. Thanks to Mark Malhotra, Shwetak Patel, Ashton Udall, and Ken Mixter for their ongoing leadership, as well as the Nest, Fit, and Assistant teams we collaborated with to build and validate Sleep Sensing on Nest Hub.

References

1. Kochanek, K. D., Murphy, S. L., Xu, J. & Arias, E. Mortality in the United States, 2013. *NCHS Data Brief* 1–8 (2014).
2. Chattu, V. K. *et al.* The Global Problem of Insufficient Sleep and Its Serious Public Health Implications. *Healthcare (Basel)* **7**, (2018).
3. Khosla, S. *et al.* Consumer Sleep Technology: An American Academy of Sleep Medicine Position Statement. *J. Clin. Sleep Med.* **14**, 877–880 (2018).
4. Rahman, T. *et al.* DoppleSleep: a contactless unobtrusive sleep sensing system using short-range Doppler radar. in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* 39–50 (Association for Computing Machinery, 2015).
5. Lien, J. *et al.* Soli: Ubiquitous Gesture Sensing with Millimeter Wave Radar. doi:10.1145/2897824.2925853.
6. Rosenberg, R. S. & Van Hout, S. The American Academy of Sleep Medicine inter-scorer reliability program: sleep stage scoring. *J. Clin. Sleep Med.* **9**, 81–87 (2013).
7. Berry, R. B. *et al. The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications : Version 2.5*. (American Academy of Sleep Medicine, 2018).
8. R Core Team. R: A Language and Environment for Statistical Computing. (2021).
9. U.S. Census Bureau. U.S. Census Bureau American Community Survey Age and Sex Table S0101. *U.S. Census Bureau* https://data.census.gov/cedsci/table (2020).
10. Ritchie, H. Age Structure. *Our World in Data* (2019).
11. Ritchie, H. Gender Ratio. *Our World in Data* (2019).
12. National Health Statistics Reports, Number 122, December 20, 2018.
13. Bastien, C. H., Vallières, A. & Morin, C. M. Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep Med.* **2**, 297–307 (2001).

14. Bogan, R. K. & Turner, J. A. New assessment tools that measure sleep vital signs: the SleepMed Insomnia Index and the Sleep Matrix. *Neuropsychiatr. Dis. Treat.* **3**, 501–510 (2007).

15. Beattie, Z. *et al.* Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals. *Physiol. Meas.* **38**, 1968–1979 (2017).

16. Cook, J. D., Prairie, M. L. & Plante, D. T. Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: A comparison against polysomnography and wrist-worn actigraphy. *J. Affect. Disord.* **217**, 299–305 (2017).

17. Kang, S.-G. *et al.* Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers. *J. Psychosom. Res.* **97**, 38–44 (2017).

18. Maskevich, S., Jumabhoy, R., Dao, P. D. M., Stout, J. C. & Drummond, S. P. A. Pilot Validation of Ambulatory Activity Monitors for Sleep Measurement in Huntington's Disease Gene Carriers. *J. Huntingtons Dis.* **6**, 249–253 (2017).

19. Meltzer, L. J., Hiruma, L. S., Avis, K., Montgomery-Downs, H. & Valentin, J. Comparison of a Commercial Accelerometer with Polysomnography and Actigraphy in Children and Adolescents. *Sleep* **38**, 1323–1330 (2015).

20. Montgomery-Downs, H. E., Insana, S. P. & Bond, J. A. Movement toward a novel activity monitoring device. *Sleep Breath.* **16**, 913–917 (2012).

21. Osterbauer, B., A. Koempel, J., L. Davidson Ward, S., M. Fisher, L. & M. Don, D. A comparison study of the Fitbit activity monitor and PSG for assessing sleep patterns and movement in children. *J. otolaryngol. adv.* **1**, 24–35 (2016).

22. Scott, H., Lovato, N. & Lack, L. The Development and Accuracy of the THIM Wearable Device for Estimating Sleep and Wakefulness. *Nat. Sci. Sleep* **13**, 39–53 (2021).

23. Cook, J. D., Eftekari, S. C., Dallmann, E., Sippy, M. & Plante, D. T. Ability of the Fitbit Alta HR to quantify and classify sleep in patients with suspected central disorders of hypersomnolence: A comparison against polysomnography. *J. Sleep Res.* **28**, e12789

(2019).

24. de Zambotti, M. *et al.* Measures of sleep and cardiac functioning during sleep using a multi-sensory commercially-available wristband in adolescents. *Physiol. Behav.* **158**, 143–149 (2016).

25. de Zambotti, M., Goldstone, A., Claudatos, S., Colrain, I. M. & Baker, F. C. A validation study of Fitbit Charge 2^{TM} compared with polysomnography in adults. *Chronobiol. Int.* **35**, 465–476 (2018).

26. Svensson, T., Chung, U.-I., Tokuno, S., Nakamura, M. & Svensson, A. K. A validation study of a consumer wearable sleep tracker compared to a portable EEG system in naturalistic conditions. *J. Psychosom. Res.* **126**, 109822 (2019).

27. Moreno-Pino, F., Porras-Segovia, A., López-Esteban, P., Artés, A. & Baca-García, E. Validation of Fitbit Charge 2 and Fitbit Alta HR Against Polysomnography for Assessing Sleep in Adults With Obstructive Sleep Apnea. *J. Clin. Sleep Med.* **15**, 1645–1653 (2019).

28. Chinoy, E. D. *et al.* Performance of Seven Consumer Sleep-Tracking Devices Compared with Polysomnography. *Sleep* (2020) doi:10.1093/sleep/zsaa291.

29. Roberts, D. M., Schade, M. M., Mathew, G. M., Gartenberg, D. & Buxton, O. M. Detecting sleep using heart rate and motion data from multisensor consumer-grade wearables, relative to wrist actigraphy and polysomnography. *Sleep* **43**, (2020).

30. Chakar, B. *et al.* Validation of midsagittal jaw movements to measure sleep in healthy adults by comparison with actigraphy and polysomnography. *Sleep Sci* **10**, 122–127 (2017).

31. de Souza, L. *et al.* Further validation of actigraphy for sleep studies. *Sleep* vol. 26 81–85 (2003).

32. Fonseca, P. *et al.* Validation of Photoplethysmography-Based Sleep Staging Compared With Polysomnography in Healthy Middle-Aged Adults. *Sleep* **40**, (2017).

33. Jean-Louis, G., Kripke, D. F., Cole, R. J., Assmus, J. D. & Langer, R. D. Sleep detection

with an accelerometer actigraph: comparisons with polysomnography. *Physiol. Behav.* **72**, 21–28 (2001).

34. Jean-Louis, G., Kripke, D. F., Mason, W. J., Elliott, J. A. & Youngstedt, S. D. Sleep estimation from wrist movement quantified by different actigraphic modalities. *J. Neurosci. Methods* **105**, 185–191 (2001).

35. Jumabhoy, R. *et al.* Validation of Consumer and Research-Grade Activity Monitors Against Polysomnography in Healthy Adults. (2019) doi:10.31234/osf.io/mx2ae.

36. Kogure, T., Shirakawa, S., Shimokawa, M. & Hosokawa, Y. Automatic sleep/wake scoring from body motion in bed: validation of a newly developed sensor placed under a mattress. *J. Physiol. Anthropol.* **30**, 103–109 (2011).

37. Kosmadopoulos, A., Sargent, C., Darwent, D., Zhou, X. & Roach, G. D. Alternatives to polysomnography (PSG): a validation of wrist actigraphy and a partial-PSG system. *Behav. Res. Methods* **46**, 1032–1041 (2014).

38. Kuo, C.-E. *et al.* Development and Evaluation of a Wearable Device for Sleep Quality Assessment. *IEEE Trans. Biomed. Eng.* **64**, 1547–1557 (2017).

39. Markwald, R. R., Bessman, S. C., Reini, S. A. & Drummond, S. P. A. Performance of a Portable Sleep Monitoring Device in Individuals with High Versus Low Sleep Efficiency. *J. Clin. Sleep Med.* **12**, 95–103 (2016).

40. O’Hare, E. *et al.* A comparison of radio-frequency biomotion sensors and actigraphy versus polysomnography for the assessment of sleep in normal subjects. *Sleep Breath.* **19**, 91–98 (2015).

41. Paquet, J., Kawinska, A. & Carrier, J. Wake detection capacity of actigraphy during sleep. *Sleep* **30**, 1362–1369 (2007).

42. Sargent, C., Lastella, M., Halson, S. L. & Roach, G. D. The validity of activity monitors for measuring sleep in elite athletes. *J. Sci. Med. Sport* **19**, 848–853 (2016).

43. Slater, J. A. *et al.* Assessing sleep using hip and wrist actigraphy. *Sleep Biol. Rhythms* **13**,

172–180 (2015).

44. Palotti, J. *et al.* Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *NPJ Digit Med* **2**, 50 (2019).

45. Tal, A., Shinar, Z., Shaki, D., Codish, S. & Goldbart, A. Validation of Contact-Free Sleep Monitoring Device with Comparison to Polysomnography. *J. Clin. Sleep Med.* **13**, 517–522 (2017).

Supplementary Materials

Supplementary Table 1

Study	**Device/Algorithm**	**N**	**Sn**	**Sp**	**Sn *t, p***	**Sp *t, p***
Consumer Wearables						
Beattie et al 2017[15]	Fitbit Surge (normal)	60	98%	35%	-1.72, 0.09	***5.01, <0.01***
Cook et al 2017[16]	Fitbit Flex (normal)	21	98%	35%	-1.62, 0.12	***4.33, <0.01***
Kang et al 2017[17]	Fitbit Flex (normal)	17	97%	36%	N/A	N/A
Maskevich et al 2017[18]	Fitbit One (normal)	7	99%	27%	-2.22, 0.07	***3.61, 0.01***
Meltzer et al 2015[19]	Fitbit Ultra (normal)	63	87%	52%	N/A	N/A
Montgomery-Downs et al 2012[20]	Fitbit Classic (normal)	24	98%	20%	-1.36, 0.19	***5.02, <0.01***
Osterbauer et al 2016[21]	Fitbit Flex (normal)	14	99%	10%	-2.81, 0.01	8.34, <0.01
Scott et al 2021[22]	Fitbit Flex (normal)	25	98%	32%	-1.51, 0.14	***4.01, <0.01***
Scott et al 2021[22]	Fitbit Alta (normal)	20	96%	39%	0.20, 0.84	***2.92, <0.01***
Beattie et al 2017[15]	Fitbit Surge (sensitive)	60	78%	80%	12.06, <0.01	-6.06, <0.01
Cook et al 2017[16]	Fitbit Flex (sensitive)	21	78%	80%	8.34, <0.01	-4.93, <0.01
Kang et al 2017[17]	Fitbit Flex (sensitive)	17	65%	82%	N/A	N/A
Meltzer et al 2015[19]	Fitbit Ultra (sensitive)	63	70%	79%	N/A	N/A
Cook et al 2019[23]	Fitbit Alta HR (sleep)	49	96%	58%	0.27, 0.79	-0.76, 0.45
deZambotti et al 2016[24]	Fitbit Charge HR (sleep)	32	97%	42%	-0.71, 0.49	***2.79, <0.01***
deZambotti et al 2018[25]	Fitbit Charge 2 (sleep)	44	96%	61%	N/A	N/A
Svensson et al 2019[26]	Fitbit Versa (sleep)	20	92%	54%	***4.39, <0.01***	0.19, 0.85
Moreno-Pino et al. 2019[27]	Fitbit Charge 2 & Alta HR (sleep)	7	89%	40%	***2.83, 0.03***	1.65, 0.15

Chinoy et al 2021[28]	Fitbit Alta HR (sleep)	49	95%	54%	N/A	N/A
Chinoy et al 2021[28]	Garmin Fenix 5S	29	99%	18%	N/A	N/A
Chinoy et al 2021[28]	Garmin Vivosmart 3	43	99%	19%	N/A	N/A
Scott et al 2021[22]	THIM	25	91%	59%	***5.27, <0.01***	-1.03, 0.31
Scott et al 2021[22]	THIM	20	89%	59%	***5.42, <0.01***	-1.07, 0.30
Roberts et al 2020[29]	Oura Ring	32	96%	41%	-0.04, 0.96	***4.69, <0.01***
		Avg	90.71%	50.05%		
		Max	99%	82%		
		Min	65%	10%		
Clinical Actigraph						
Chakar et al 2017[30]	Actiwatch 2 (default)	38	96%	48%	N/A	N/A
De Souza et al 2003[31]	Mini motionlogger (Cole-Kripke)	21	99%	34%	N/A	N/A
De Souza et al 2003[31]	Mini motionlogger (Sadeh)	21	97%	44%	N/A	N/A
Fonseca et al 2017[32]	Actiwatch Spectrum (default)	49	97%	46%	-0.83, 0.41	2.04, 0.05
Jean-Louis et al 2001[33]	Actillume I	31	95%	31%	N/A	N/A
Jean-Louis et al 2001[34]	Actillume I	5	99%	28%	N/A	N/A
Jumabhoy et al 2019[35]	Actiwatch 2 (default)	22	97%	27%	N/A	N/A
Kogure et al 2011[36]	Micro-Mini & Mini Motionlogger	6	99%	34%	N/A	N/A
Kosmadopoulos et al 2014[37]	Actiwatch-64 (medium sensitivity)	22	96%	38%	0.42, 0.68	***3.61, <0.01***
Kuo et al 2017[38]	Non-commercial	59	95%	71%	N/A	N/A
Markwald et al 2016[39]	Actiwatch-64 (medium sensitivity)	26	97%	37%	N/A	N/A
Montgomery-Downs et al 2012[20]	Actiwatch-64	24	96%	39%	0.42, 0.68	***2.14, 0.04***
O'Hare et al 2015[40]	Actiwatch	20	97%	34%	-0.96, 0.35	***4.57, <0.01***
Paquet et al 2007[41]	Actiwatch-L	15	95%	54%	0.82, 0.43	0.06, 0.95
Roberts et al 2020[29]	ActiGraph Link (Cole-Kripke)	32	94%	57%	2.03, 0.05	-0.46, 0.65

Roberts et al 2020[29]	ActiGraph Link (Sadeh)	32	91%	65%	3.40, <0.01	-2.19, 0.04
Roberts et al 2020[29]	Actiwatch Spectrum Plus	32	98%	37%	-1.94, 0.06	***4.22, <0.01***
Sargent et al 2016[42]	Actiware	16	88%	77%	N/A	N/A
Slater et al 2015[43]	GTX3+	108	90%	46%	***8.33, <0.01***	***4.65, <0.01***
Scott et al 2021[22]	Actiwatch 2	25	95%	35%	1.59, 0.13	***5.47, <0.01***
Scott et al 2021[22]	Actiwatch 2	20	95%	59%	1.42, 0.17	-0.97, 0.35
Chinoy et al 2021[28]	Actiwatch 2 (medium sensitivity)	98	97%	39%	N/A	N/A
Palotti et al 2019[44]	Actiwatch Spectrum (multiple algorithms)	2237	93%	50%	***24.15, <0.01***	***23.10, <0.01***
		Avg	93.38%	48.69%		
		Max	99%	77%		
		Min	88%	27%		
Under-mattress sensor						
Tal et al 2017[45]	EarlySense	85	93%	80%	N/A	N/A
Chinoy et al 2021[28]	EarlySense Live	51	96%	47%	N/A	N/A
		Avg	93.81%	67.88%		
		Max	96%	80%		
		Min	93%	47%		
Radar						
Chinoy et al 2021[28]	ResMed S+	51	93%	51%	N/A	N/A
Chinoy et al 2021[28]	SleepScore Max	42	94%	50%	N/A	N/A
O'Hare et al 2015[40]	SleepMinder	20	95%	42%	1.02, 0.32	***2.89, <0.01***
O'Hare et al 2015[40]	SleepDesign	20	96%	38%	-0.06, 0.96	***5.38, <0.01***
		Avg	94.17%	47.38%		
		Max	96%	51%		
		Min	93%	38%		

Aggregate of all modalities (weighted average & standard deviation)						
			92.91±4.70%	49.56±11.52%	3.44, <0.01	1.44, 0.16
		Max	99%	82%		
		Min	65%	10%		
		Algorithm superior N (%)			6 (20.00%)	18 (60.00%)

Supplementary Table 1. Existing studies validating various devices against polysomnography. Reported mean sensitivity and specificity from various consumer and clinical devices along with category-specific, weighted averages, maxima, and minima. *t*-statistics and *p*-values provided for comparisons between algorithm performance and previously published means, for those data that had reported a measure of dispersion (e.g., standard deviation). While these are not head-to-head comparisons, a positive *t*-statistic suggests superior performance of the algorithm; otherwise, there is an “N/A” when no measures of dispersion were available to allow for statistical comparisons. All instances where the algorithm was statistically superior in one domain (e.g., sensitivity) while not statistically worse in the other (e.g., specificity) in comparison to the reported performance are bolded and italicized, with a count and percentage of all such instances provided at the bottom of the table. Comparison was also made against an aggregate performance of all devices, roughly approximated by weighted mean of means and standard deviation of means. Abbreviations: Sn - sensitivity; Sp - specificity.