Nature and Science of Sleep

Elevate Series

Open Access Full Text Article

ORIGINAL RESEARCH

Quantification of Differences in Sleep Measurement by a Wrist-Worn Consumer Wearable Compared to Research-Grade Accelerometry and Sleep Diaries of Female Adults in Free-Living Conditions

Cindy R Hu[1], Caitlin Delaney[2], Jorge E Chavarro[2–4], Francine Laden[1–3], Rachel Librett[4], Laura Katuska[4], Emily R Kaplan[5], Li Yi[4,6], Michael Rueschman[5], Joe Kossowsky[7], Jukka-Pekka Onnela[8], Brent A Coull[8], Susan Redline[3,5], Peter James[1,6,9], Jaime E Hart[1,2]

[1]Department of Environmental Health, Harvard T.H. Chan School of Public Health; Boston, Boston, MA, USA; [2]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School; Boston, Boston, MA, USA; [3]Department of Epidemiology, Harvard T.H. Chan School of Public Health; Boston, Boston, MA, USA; [4]Department of Nutrition, Harvard T.H. Chan School of Public Health; Boston, Boston, MA, USA; [5]Division of Sleep of Sleep and Circadian Disorders, Departments of Medicine and Neurology, Brigham and Women's Hospital and Harvard Medical School; Boston, Boston, MA, USA; [6]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute; Boston, Boston, MA, USA; [7]Department of Anesthesiology, Critical Care, and Pain Medicine, Boston Children's Hospital and Harvard Medical School; Boston, Boston, MA, USA; [8]Department of Biostatistics, Harvard T.H. Chan School of Public Health; Boston, Boston, MA, USA; [9]Department of Public Health Sciences, University of California, Davis School of Medicine; Davis, Davis, CA, USA

Correspondence: Cindy R Hu, Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Avenue, Boston, MA, 02115, USA, Email cindyhu@g.harvard.edu

Purpose: The objective of this study is to compare sleep measurements by a consumer-wearable with research-standard actigraphy coupled with sleep diaries in free-living female adults.

Methods: Forty-seven females in the Nurses' Health Study 3 (NHS3) participated in the Sleep and Physical Activity Validation Substudy (SPAVS), where they were asked to concurrently wear a consumer wearable (Fitbit Charge, Models 3 or 5) and a research-grade accelerometer (Actigraph, GT3X+ or Actisleep) on the same wrist and fill out a smartphone-based sleep diary for fourteen consecutive days. We compared measures of total sleep time (TST), time in bed (TIB), and sleep efficiency (SE) from the consumer wearable with actigraphy measures as our research-standard reference for TST and SE and self-reported sleep diary as our reference for TIB. We calculated mean absolute percent error (MAPE) and intra-class correlations (ICC), as well as Bland-Altman analyses to compute mean difference and limits of agreement.

Results: For all three measures, the consumer wearable underestimated sleep parameters relative to research-standard actigraphy, with a mean bias of −16.0 minutes and −11.2 minutes for TST and TIB, respectively, and −1.0% for SE. In terms of agreement, TST (MAPE = 11.18%; ICC = 0.79) and TIB (MAPE = 10.45%; ICC = 0.74) had similar MAPES and ICCs, while and SE (MAPE = 5.09%; ICC = 0.39) had a lower ICC.

Conclusion: In the NHS3 SPAVS, the wearable sleep measurements modestly underestimated wrist actigraphy measures of TST, TIB, and SE from sleep over multiple days; within sleep measures assessed, TST and TIB had greater agreement with research-grade accelerometry than SE.

Keywords: wearables, fitbit, sleep, actigraphy, accelerometer, women

Introduction

Sleep is an essential process of restoration for the human body.[1,2] Due to its links to outcomes across mental, physical, and chronic disease dimensions of health, human sleep is a subject of epidemiologic interest and importance.[3]

The value of population-level insights on sleep health underscores the importance of sleep measurement tools that are robust and feasible in large studies under non-laboratory conditions with long-term follow-up. Epidemiological studies of sleep often rely on self-reported measures. While self-reported measures are easier to integrate into large studies, device-based measures of sleep are valuable because self-reported measures of duration and quality have been shown to differ from

Received: 4 April 2025
Accepted: 9 August 2025
Published: 27 August 2025

objective sleep measures. In a study by Lauderdale et al, sleep duration was consistently overestimated in self-reported measures, with the difference between the two measures widening as sleep duration decreased.[4–6] The gold-standard of sleep measurement, polysomnography (PSG), is complex and requires expertise and specialized equipment. This typically limits its application to a single night and is not only expensive but also does not capture typical at-home sleep for the average person.[7] Research-grade actigraphy devices that are worn on a participant's wrist are often used, which have been shown to yield valid measures of sleep relative to PSG.[8,9] However, research-grade objective sleep measurement tools like actigraphy watches are not well-suited to long-term individual-use and can be cumbersome to wear.

The advent of mobile health tools and commercially available wearable devices presents a potentially cost-effective, more accessible option for objective sleep measurement of large or geographically dispersed study populations, as well as enabling longer-term monitoring in patients.[10] Consumer-wearable sleep trackers can provide new avenues of sleep health assessment for patients in clinical settings.[11] Like in research settings, sleep medicine tools like PSG have constrained periods of possible measurement, while consumer-wearables offer possibilities of sleep monitoring over longer windows that can be useful for chronic disease prevention. In some settings, providing patients with consumer wearables can be a part of care, as it gives patients a way to track their own health behaviors.[12] Additionally, tracking sleep over longer windows of time can be valuable for the promotion of health in specific populations such as women, who experience cyclical and lifetime hormonal changes that impact sleep.[13]

The development of mobile health technologies like wrist-worn sleep and activity monitors provides an opportunity to quantitatively assess sleep in ecologically valid and free-living (ie, non-laboratory) conditions on a larger scale and over longer time periods.[14,15] There has been a marked increase in use of commercial-grade wearables in NIH-sponsored research, including the All of Us Research Program and Risk Underlying Rural Areas Longitudinal Heart and Lung Study.[16] However, their novelty and the proprietary nature of their product development and scoring algorithms raise questions about their interchangeability with research-grade accelerometers.[17]

In previous validation studies of mobile health sleep measurement tools, consumer wearables have demonstrated high levels of accuracy and sensitivity for correctly identifying sleep episodes compared to PSG but have been studied only in laboratory settings rather than free-living conditions that conform to everyday sleep.[18–23] In this study, we aim to quantify differences in sleep measurement by a consumer wearable and research-grade accelerometry in the healthy free-living female US-based adult population of the Nurses' Health Study 3 Sleep and Physical Activity Validation Substudy.

Materials and Methods

Nurses' Health Study 3 Sleep and Physical Activity Validation Substudy

The Nurses' Health Study 3 (NHS3) is a prospective cohort of nurses and nursing students in the United States and Canada that began enrollment in 2010. Participants in NHS3 were born January 1st, 1965 or later and fill out internet-based surveys on health and health behaviors approximately every six months. Within NHS3, the Sleep and Physical Activity Validation Substudy (SPAVS) was conducted with a goal of enrolling 50 female NHS3 participants, with data collection beginning in September 2023 and concluding in May 2024. In SPAVS, participants were asked to wear a research-grade accelerometer and a consumer wearable device concurrently on their wrist for fourteen consecutive days and to complete daily sleep logs on their smartphone using a research application called Beiwe.[24] The eligibility criteria for participation in NHS3 SPAVS was as follows: owned a smartphone, lived in the contiguous United States, had reliable access to Wi-Fi at least once a week, did not work night shifts during the collection period, had ability to wear wrist-worn devices 24-hours a day for 14 days, did not use sleep medications or have a sleep disorder, did not use a continuous positive airway pressure (CPAP) machine, was not pregnant, and had no health conditions that result in a tremor. For their participation in the study, participants were able to keep the consumer wearable. We sent study kits to 63 participants. We excluded participants with no days with simultaneous research-grade accelerometry, sleep diary, and consumer wearable sleep data (N = 12); this was due to device (ie, research-grade accelerometer, consumer wearable, or smartphone application) issues (N = 9), participant non-adherence with the protocol (eg, not wearing the research-grade accelerometer and consumer wearable on the same wrist) (N = 2), or never returning the research-grade accelerometry device (N = 1) (Figure 1). Our final analytic sample size comprised 47 participants.

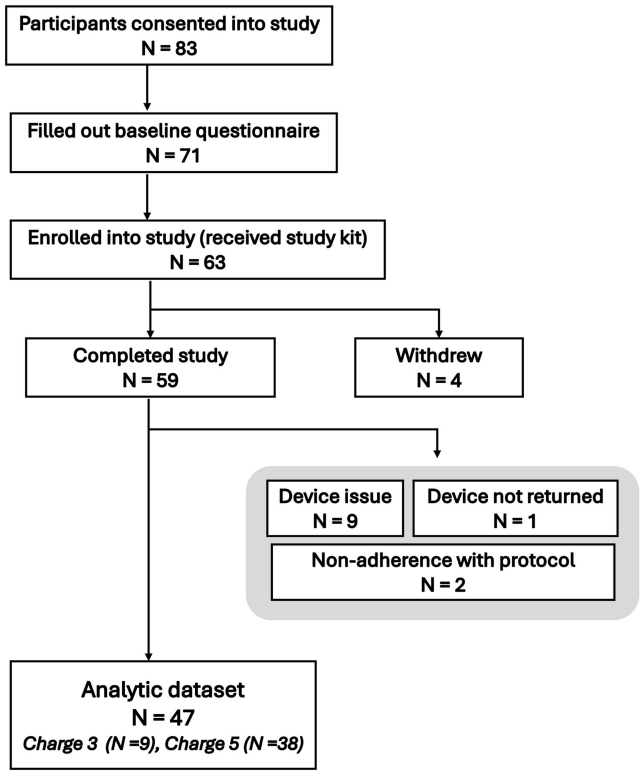

Figure 1 Recruitment, data collection and derivation of the analytic dataset of 47 in the Nurses' Health Study 3 (NHS3) Sleep and Physical Activity Validation Substudy (SPAVS).

Sleep Questionnaires and Sleep Diary

After completing an online screener and completing an online consent form, participants in SPAVS completed a baseline web questionnaire that included the Pittsburgh Sleep Quality Index (PSQI) and the reduced version of the Morningness-Eveningness Questionnaire (rMEQ); the PSQI and rMEQ questionnaires and scoring are accessible in their introductory publications by Buysse et al and Adan and Almirall, respectively.[25,26] At the beginning of the device data collection period, participants were asked to download the research smartphone application, which served as the platform to complete daily surveys. During the study period, participants completed an electronic sleep diary every morning, reporting their bedtime from the previous night and waketime. Participants also reported any naps taken in the previous day. Sleep diaries are often collected with research actigraphy, to assist with annotation. On the seventh day of the study period, participants filled out the Patient-Reported Outcomes Measurement Information System (PROMIS) Sleep-Related Impairment scale (Version 8a), whose contents are accessible in the publication by Yu et al, on the smartphone application; their scores were tabulated based on the cut-points provided by the Health Measures scoring service with the calibration sample set to default PROMIS Sleep Wave 1 (accessible from PROMIS Health Organization via https://www.assessmentcenter.net/ac_scoringservice).[27] As a backup in case, there were any issues with the smartphone application, participants were able to document their sleep periods on a provided paper sleep diary.

Research-Grade Accelerometer

Participants received wrist-worn research-grade accelerometry devices to wear on their non-dominant wrist for all 14 days of the study. We used GT3X+ and Actisleep models from Actigraph (Pensacola, FL) that were configured for 60-second epochs, with a sampling rate of 30 Hertz. The Actigraph GT3X+ and Actisleep are both models of actigraphy watch that have a tri-axial accelerometer; they primarily differ in additional research features (ie, light exposure sensor on the Actisleep) that were not utilized in this study. The actigraphy data were downloaded and scored using Actilife software (Actilife 6; Version 6.13.4); sleep periods were scored blinded to consumer wearable data. Sleep metrics were

generated by inputting the start and end times of sleep periods from the sleep diaries into the scoring software, applying the Cole-Kripke algorithm, and exporting the resulting sleep period-level summary measures.[28]

Consumer Wearable Device

Participants wore a Fitbit (San Francisco, CA) alongside the research-grade accelerometer on their non-dominant wrist for the study period with consistent wrist placement for all participants (Figure S1). Due to the device availability, both Charge 3 and Charge 5 models were used; among the 47 participants in the analysis, 9 participants wore a Charge 3, and 38 participants wore a Charge 5. "According to manufacturer, there are some differences between Charge 3 and Charge 5 devices; the Charge 5 has additional GPS capabilities, and electrodermal activity sensors, which measure sweat levels on skin to assess stress levels. However, no explicitly advertised differences in hardware or software for sleep estimation. Compared to the Charge 3, the Charge 5 has some additional sleep features for users, such as a sleep score metric and sleep mode; however, no use of sleep mode was included in our study protocol, and no sleep score measures were included in our analysis.[29,30] Sleep data were synced to and downloaded from Fitabase. Fitabase is a platform that supports the collection and management of data from their consumer wearable devices. We used Fitabase to monitor participants' data on a daily basis and download their sleep data.

Statistical Analysis

All statistical analyses were conducted in R (Version 4.4.0; Vienna, Austria). Our inclusion criteria for sleep periods in this analysis were having concurrent valid research-grade accelerometer and consumer wearable measures for a given sleep period. We included the main sleep periods and naps in this analysis. We conducted sleep period-level comparisons for three sleep metrics: total sleep time (TST), time in bed (TIB), and sleep efficiency (SE). For TST, we compared the research-grade accelerometer's measure of TST to the consumer wearable's provided minutesAsleep measure. For TIB, we used the time elapsed from the bedtime and waketime reported in the sleep diary and the consumer wearable's TimeInBed metric, for the research-grade accelerometer and the consumer-grade device, respectively. For SE, we used the research-grade accelerometer's SE measure (TST*100/TIB) and calculated SE for the consumer wearable using their derived minutesAsleep and TimeInBed measures using the following formula: minutesAsleep*100/TimeInBed. For all three sleep outcomes, we calculated mean absolute percent error (MAPE) for each sleep period; the MAPE was calculated as the average of the absolute values of: (consumer wearable's measure – research-grade accelerometer's measure)*100/research-grade accelerometer's measure. We also calculated intra-class correlations (ICCs), for which we specified a random intercept for participant due to the within-participant repeated measures of the study and an autoregressive correlation structure to account for correlation in adjacent days of observations. ICCs served as our sole measure that accounts for clustering attributable to longitudinal measurements by participant and consecutive days of observation. We used Bland-Altman analyses to calculate limits of agreement (LOA) for each measure, specifying a 0.95 significance level; thus, LOAs denote the range in which 95% of differences lie. The Bland-Altman analyses were calculated so that a positive mean difference value signified overestimation by the consumer wearable, and negative mean differences indicated consumer wearable underestimation. These differences in sleep metrics were also visualized using Bland-Altman plots.[31] For the primary analyses, we pooled observations of the two models of consumer wearable (Charge 3 and Charge 5) but also conducted secondary analyses stratified by device to assess if data from a particular device model were driving our observed results. We also conducted stratified analyses by PSQI, rMEQ, and PROMIS sleep-related impairment score.

Results

SPAVS participants were female and predominantly white, with an average age of 40.4 years (SD = 6.9) (Table 1). There were no differences in demographic characteristics of participants who were and were not included in the analysis (Table S1). The average number of sleep periods contributed during the 14-day study period was 12.5, with a sleep duration and sleep efficiency of 7.2 hours and 90.2%, respectively, as measured by the research-grade accelerometer (Table 2). For the consumer wearable, the average sleep duration was 6.9 hours and average efficiency was 89.1%. There were no major differences

Table 1 Demographic Characteristics of 47 Female Participants in the Nurses' Health Study 3 Sleep and Physical Activity Validation Substudy

	N = 47
Age (Mean, SD)	40.4 (6.8)
Race (N, %)	
White	45 (95.7)
Asian	2 (4.3)
Marital Status (N, %)	
Ever married	32 (68.1)
Never married	15 (31.9)
Parity (N, %)	
Nulliparous	2 (4.3)
Parous	45 (95.7)
PSQI Score	
Mean (SD)	7.5 (4.0)
Score > 5 (N, %)	27 (57.4)
Missing	2 (4.3)
rMEQ (N, %)	
Morning	16 (34.0)
Neither	27 (57.4)
Evening	3 (6.4)
Missing	1 (2.1)
PROMIS SRI8a	
Mean (SD)	52.5 (5.3)
Score < 50 (N, %)	18 (38.3)
Missing (N, %)	5 (10.6)
Average number of sleep periods (n)	12.5 (3.6)

Abbreviations: PSQI, Pittsburgh Sleep Quality Index (PSQI); rMEQ, Reduced Morningness-Eveningness Questionnaire; PROMIS SRI8a, Patient-Reported Outcomes Measurement Information System Battery for Sleep-related Impairment, Version 8a.

Table 2 Sleep Period Characteristics of 47 Female Participants in the Nurses' Health Study 3 Sleep and Physical Activity Validation Substudy

	Total Sleep Time (hours)	Time in Bed (hours)	Sleep Efficiency (%)
Research-grade accelerometer			
Mean (SD)	7.2 (1.6)	7.9 (1.8)	90.1 (6.4)
Median (25^{th}, 75^{th})	7.3 (6.6, 8.1)	8.0 (7.4, 8.9)	91.3 (87.5, 94.2)
Consumer wearable			
Mean (SD)	6.9 (1.7)	7.7 (1.9)	89.1 (3.4)
Median (25^{th}, 75^{th})	7.1 (6.3, 7.8)	8.0 (7.2, 8.8)	89.2 (87.3, 91.1)
Difference between consumer wearable and research-grade accelerometer			
Mean (SD)	−0.3 (1.1)	−0.2 (1.3)	−1.0 (5.9)
Median (25^{th}, 75^{th})	−0.2 (−0.5, 0.2)	−0.0 (−0.4, 0.3)	−1.9 (−4.6, 1.5)

between models of consumer wearable used in the study for demographics of participants in the study (Table S2), nor average TST, TIB, and SE (Table S3).

The MAPE and ICCs for consumer wearable compared to research-grade accelerometer and sleep diary measures are presented in Table 3. The MAPE for TST was 11.18%, with an ICC of 0.79. The MAPE and ICC for TIB were of similar

https://doi.org/10.2147/NSS.S530812

Table 3 Mean Absolute Percent Error and Intraclass Correlation for Total Sleep Time, Total Time in Bed, and Sleep Efficiency Overall and by Sleep Characteristics in the Nurses' Health Study 3 Sleep and Physical Activity Validation Substudy

	All Participants	PSQI Score		rMEQ Score		PROMIS SRI T-score	
	N = 47	PSQI ≤ 5 N = 27	PSQI > 5 N = 18	Morning Type N = 16	Not Morning Type N = 30	PROMIS < 50 N = 24	PROMIS ≥ 50 N = 18
Total Sleep Time							
MAPE (%)	11.18	9.10	13.1	13.31	10.21	12.84	8.90
ICC	0.79	0.71	0.91	0.79	0.79	0.76	0.85
Total Time in Bed							
MAPE (%)	10.45	7.10	13.28	12.51	9.47	12.86	7.2
ICC	0.74	0.64	0.90	0.76	0.73	0.69	0.85
Sleep Efficiency							
MAPE (%)	5.09	5.18	5.15	4.95	5.20	5.30	5.01
ICC	0.39	0.41	0.34	0.52	0.32	0.43	0.28

Abbreviations: MAPE, Mean absolute percent error; ICC, Intraclass correlation; PSQI, Pittsburgh Sleep Quality Index (PSQI); rMEQ, Reduced Morningness-Eveningness Questionnaire; PROMIS Battery, Patient-Reported Outcomes Measurement Information System Battery for Sleep-related Impairment.

magnitude, at 10.45% and 0.75, respectively. For SE, the MAPE was 5.09%, and the ICC was 0.39. In stratified analyses, we did not observe consistent patterns of differential agreement across strata of higher/lower PSQI scores, chronotype, or higher/lower PROMIS score for MAPEs or ICCs (Table 3). In sensitivity analyses by model of consumer wearable, we did not observe meaningful differences in MAPEs and ICCs (Table S4).

From the Bland-Altman analysis (Figure 2), we observed that the consumer wearable underestimated TST, TIB, and SE. TST had a mean difference of −16.0 minutes (95% LOAs: −143.2, 111.3). Similarly, in TIB, the mean difference was −11.2 minutes (95% LOAs: −167.0, 144.7 minutes). For SE, the mean difference of −1.0% (95% LOAs: −12.6%, 10.7%) indicated underestimation by the consumer wearable relative to research-grade actigraphy. In our Bland-Altman plots, we observed no clear visual differences in spread between the two models of consumer wearable included in the study (Figure S2). We did see a pattern of smaller biases for the Charge 5 device compared to the Charge 3 device (Figure S2). However, due to the smaller sample size of Charge 3 device users (N = 9), we have limited power to interpret this result as indicative of differential performance by device model.

Discussion

In this study, we compared measurements of TST, TIB, and SE from a consumer wearable with research-standard wrist actigraphy combined with sleep diaries in a population of healthy female adults. Overall, we observed high agreement between TST, TIB, and SE measurements from the consumer wearable and our research standard measures of actigraphy and daily self-report sleep diary, as well as a modest underestimating bias by the consumer wearable for all sleep outcomes.

Our findings of relatively small biases in TST and SE parallel prior studies examining sleep measures of similar consumer wearable models, such as products from the same company. Within NHS3 SPAVS and relative to wrist actigraphy, the consumer wearable underestimated the sleep metrics we investigated: TST by 16 mins and just short of 1% for SE. However, the calculated LOAs with 95% significance for all three outcomes included zero. The magnitude of differences we observed was comparable to other studies (ie, within a range of less than 20 minutes for TST, and less than 5% for SE and LOAs that similarly encompass zero), though there are mixed findings on whether this model series of consumer wearable devices overestimate or underestimate research standard measures. Some prior studies using PSG and using actigraphy also reported that consumer wearables of this series model underestimated the reference measure, with biases ranging from approximately 7 to 14 minutes for TST.[19–21,32,33] However, some studies comparing other

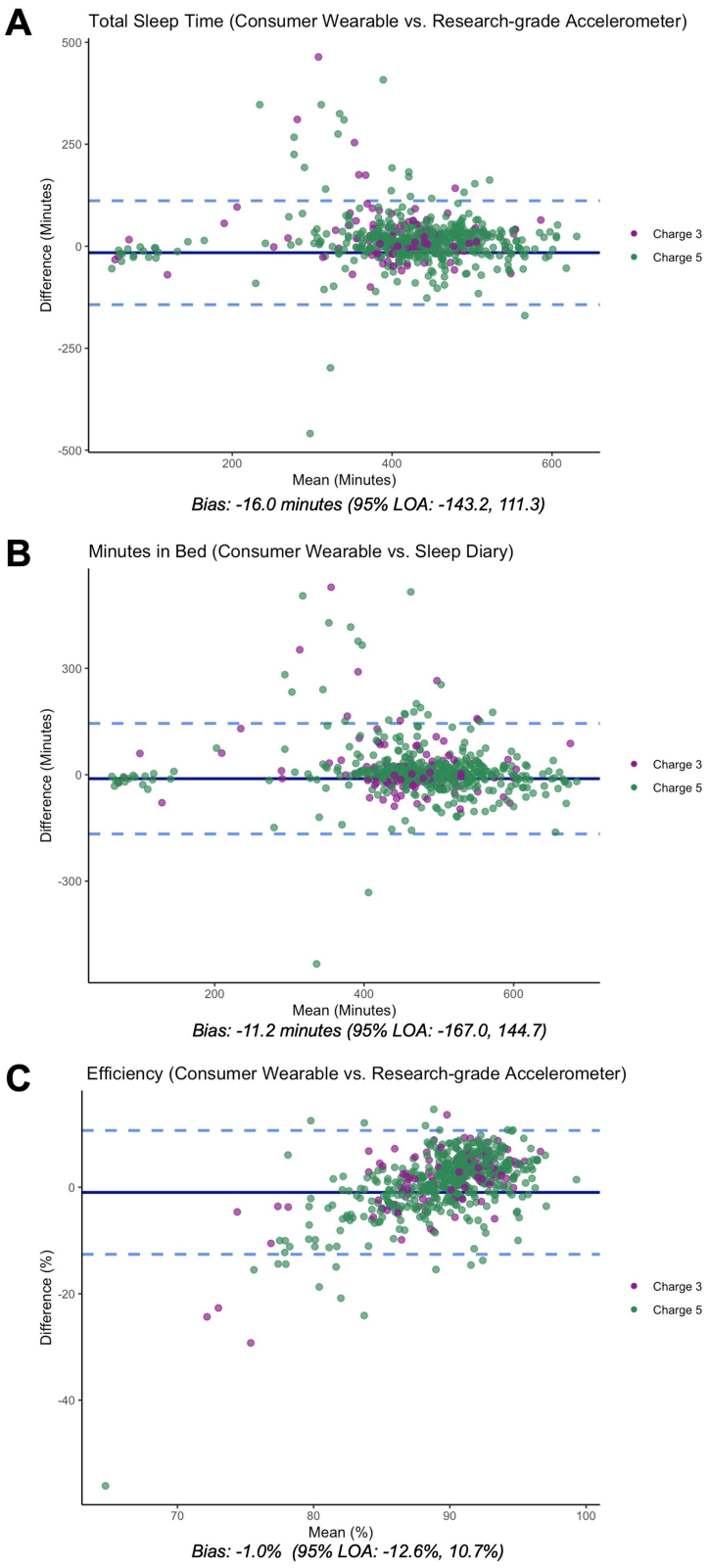

Figure 2 Bland-Altman plots for total sleep time in Panel (**A**), total time in bed in Panel (**B**), and sleep efficiency in Panel (**C**) in the Nurses' Health Study 3 (NHS3) Sleep and Physical Activity Validation Substudy (SPAVS).

https://doi.org/10.2147/NSS.S530812

models of this consumer wearable to PSG and EEG reported the consumer-grade device overestimating TST, with a similar magnitude of bias (9 to 17 minutes).[18,22] Compared to TST, SE is understudied; our study contributes to the extant three studies, for which the SEs range from −4% to 4%.[20,22,33] Our study of healthy adult females supports the literature that suggests that there are no consistent patterns in the direction of bias nor meaningful differences in magnitude across types of study populations (eg, healthy adults, Huntington disease patients, adolescents), device models, or reference measures (ie, PSG or actigraphy).[18,20,22,32,33] Several previous studies with models of this consumer wearable were in laboratory settings, and our findings from a field-based study design provide the insight that the magnitude of differences observed is like that of lab studies. Our study also contributes to the limited literature on TIB for this model series of consumer wearable. A study of consumer wearables from the same company compared with wrist actigraphy by McMahon et al found that the consumer-grade device underestimated TST and TIB by 71 and 77 minutes, respectively; however, this study had participants wear research-grade actigraphy watches and consumer wearable devices on opposite wrists, which may contribute to the observed differences in bias.[34]

We also conducted stratified analyses to assess if there were differences in consumer wearable agreement between participants with different status across dimensions of sleep health. Within NHS3 SPAVS, we observed elevated PSQI and PROMIS scores in a majority (57% and 62% for PSQI and PROMIS, respectively) of participants; while we may describe this population as "healthy" regarding lack of sleep disorder diagnoses, these scores indicate poor sleep or sleep-related impairment. Based on the PSQI and PROMIS metrics, we saw that participants with poorer sleep quality and participants with greater sleep-related impairment had greater agreement in TST and less agreement in SE. From our stratified analyses, there did not appear to be consistent differential performance for "better" sleepers within a population of healthy adults. The consumer wearable's consistency in performance by study population is additionally supported by the parity of findings for this model series of consumer wearables in sleep-disordered and sleep-impacted populations.[19,20]

For the time-based sleep measures like TST and TIB, we observed a small number of observations (<4%) with meaningful discrepancies (ie, over 3 hours) between consumer wearable and the reference measure (research-grade accelerometry for TST, self-report diary for TIB), which we kept in the analysis; among these observations, there was an even distribution of observations in which the consumer wearable was overestimating and was underestimating. When we investigated these discrepant observations via visual inspection, we saw that the timing of onset of both TIB and TST periods agreed very well. These factors suggested that the discrepancies in these particular observations comes from when a consumer wearable device is either premature or delayed in detecting the end of a sleep period.

Across all three outcomes, although the LOAs include zero, their ranges are relatively wide, reflecting variation among the calculated biases of individual sleep observations. The wide LOAs suggest reduced precision, but the modest mean biases indicate that there is not a strong systematic bias, which reflects accuracy. The lower precision of the consumer wearable would introduce non-differential measurement error. In regards to use in clinical and epidemiologic applications, the implications of this measurement error depend on whether sleep is used as an exposure or outcome; for analyses with sleep as an exposure, potential effects would be attenuated towards the null, while for analyses with sleep as an outcome, there would be no introduced bias.

One of the central limitations of this study is intrinsic to studies using consumer wearable devices. Our findings are generalizable to observations from this consumer wearable model with the algorithms and firmware active at the time of data collection; since the consumer wearable's algorithms and firmware are proprietary, we cannot definitively know if or when these software or firmware features were or may be updated in ways that impact the generation of sleep metrics. Another limitation of this study is that NHS3 SPAVS was an all-female study population, so we are unable to assess sleep-related sex differences in consumer wearable measurement. Additionally, our study sample is very racially homogenous, with a predominance of White participants. With the limited demographic diversity of our study sample, we are unable to assess how the performance of the consumer wearable relative to research-grade accelerometry may vary in other populations. Due to the proprietary nature of the sleep estimation process, it is difficult to speculate on whether certain behaviors or characteristics like skin tone may impact the accuracy of sleep measurement.

The strengths of this study lie within the study design. First, we were able to study individuals in free-living conditions, providing insights applicable to future epidemiologic studies of sleep that have non-laboratory conditions. Second, we were able to conduct our studies in a group of nearly 50 participants, with 14 days of longitudinal measurements, which augmented

the power of our findings. Relative to previous studies, this study design has increased power from a larger sample size (N = 47) and longer study period (14 consecutive days); prior studies comparing sleep measures from research-grade accelerometers and consumer wearables of free-living adults, the study sample size has been approximately 20 participants who were followed for 3 days.[35,36]

Consumer wearables often provide measures of sleep staging, in addition to the metrics evaluated in this study. Rather than using research-grade actigraphy in free-living conditions as a reference, sleep staging measures are compared to PSG measures in laboratory studies. Sleep staging measures from this model series of consumer wearable have been compared to PSG in several smaller scale studies that report mixed results indicating moderate accuracy.[18,19,33,37] Since we rely on the insights of studying a larger group of individuals in conditions representative of daily life to determine a sleep metric's robustness, it can be more challenging to evaluate the suitability for use in large scale studies for these sleep staging measures.

Consumer wearables represent exciting opportunities to advance epidemiologic studies of sleep and expand assessment in sleep medicine but require careful consideration for integration into research settings.[38] Ideally, the process of evaluating consumer wearables for research comprises three stages: first, a laboratory-based study with comparison to PSG, then, a field-based study with comparison to a validated measure (ie, actigraphy), and finally, studies of specific subpopulations.[39] In context of the several extant studies comparing models of this consumer wearable to PSG in lab settings, this study contributes to the advancement of the second stage by characterizing differences from research standard measures in non-laboratory conditions in a well-defined sample.

Conclusion

The emergence of mobile health wearable technologies like wrist-worn sleep and physical activity tracking devices has implications for increasing feasibility of sleep data collection; these data can be used in population-level sleep studies or in clinical settings to assist in monitoring patient sleep health. Our comparison of a consumer wearable and wrist-worn actigraphy among free-living healthy adult females demonstrates that generally high agreement is retained in non-laboratory conditions. Sleep insights into a female population can be valuable for advancing women's sleep health; women's sleep has specific social and biological considerations, such as social obligations from gender roles for women that shape their sleep opportunities as well as biological considerations like hormonal cycles that impact sleep physiology for females. More broadly, improving our understanding of how consumer wearable sleep monitors in free-living conditions facilitates their implementation into sleep medicine as well as into population-level studies, which are becoming an integral part of future investigations of the links of various epidemiologic exposures including environmental, contextual, and psychosocial factors with sleep.

Abbreviations

ICC, Intraclass correlation; MAPE, Mean absolute percent error; NHS3, Nurses' Health Study 3; PROMIS SRI8a, Patient-Reported Outcomes Measurement Information System Battery for Sleep-related Impairment, Version 8a; PSG, Polysomnography; PSQI, Pittsburgh Sleep Quality Index; rMEQ, Reduced morningness-eveningness questionnaire; SE, Sleep efficiency; SPAVS, Sleep and Physical Activity Validation Substudy; TIB, Time in bed; TST, Total sleep time.

Data Sharing Statement

The data used in our analyses are not publicly available due to privacy reasons. Researchers interested in obtaining access to NHS3 data and computing code should submit an external collaborator form (https://www.nhs3.org/for-researchers/).

Ethical Approval and Informed Consent

The study protocol was approved by the institutional review boards (IRBs) of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health. All participants provided informed consent to participate in the study. The Declaration of Helsinki was followed in the study.

Acknowledgments

We would like to thank the staff and participants of the Nurses' Health Study 3, without whom this study would not have been possible.

Author Contributions

Cindy R. Hu: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft
Caitlin Delaney: Project administration, Investigation, Writing – review and editing
Jorge E. Chavarro: Conceptualization, Funding acquisition, Supervision, Writing – review and editing,
Francine Laden: Data curation, Writing – review and editing
Rachel Librett: Project administration, Investigation, Writing – review and editing
Laura Katuska: Project administration, Investigation, Writing – review and editing
Emily R. Kaplan: Data curation, Writing – review and editing
Li Yi: Data curation, Writing – review and editing
Michael Rueschman: Data curation, Writing – review and editing
Joe Kossowsky: Data curation, Writing – review and editing
J.P. Onnela: Project administration, Resources, Writing – review and editing
Brent A. Coull: Methodology, Writing – review and editing
Susan Redline: Resources, Methodology, Writing – review and editing
Peter James: Conceptualization, Resources, Writing – review and editing
Jaime E. Hart: Conceptualization, Supervision, Writing – review and editing
All authors agreed on the journal to which the article was submitted; agreed on version accepted for publication and agree to take responsibility and be accountable for the contents of the article.

Funding

Funding for the cohort and these analyses came from National Institutes of Health grants U01 HL145386, R24 ES028521, P30 ES000002, and F31 ES035252.

Disclosure

Brent Coull reports grants from Apple, Inc. The other authors have no conflicts of interest to disclose.

References

1. Buysse DJ. Sleep health: can we define it? does it matter? *Sleep*. 2014;37(1):9–17. doi:10.5665/sleep.3298
2. Zee PC, Turek FW. Sleep and health: everywhere and in both directions. *Arch Internal Med*. 2006;166(16):1686–1688. doi:10.1001/archinte.166.16.1686
3. Ferrie JE, Kumari M, Salo P, Singh-Manoux A, Kivimaki M. Sleep epidemiology--a rapidly growing field. *Int J Epidemiol*. 2011;40(6):1431–1437. doi:10.1093/ije/dyr203
4. Jackson CL, Patel SR, Jackson WB, Lutsey PL, Redline S. Agreement between self-reported and objectively measured sleep duration among white, black, hispanic, and Chinese adults in The United States: multi-ethnic study of atherosclerosis. *Sleep*. 2018;41(6):zsy057. doi:10.1093/sleep/zsy057
5. Jackson CL, Ward JB, Johnson DA, Sims M, Wilson J, Redline S. Concordance between self-reported and actigraphy-assessed sleep duration among African-American adults: findings from the Jackson Heart Sleep Study. *Sleep*. 2020;43(3):zsz246. doi:10.1093/sleep/zsz246
6. Lauderdale DS, Knutson KL, Yan LL, Liu K, Rathouz PJ. Self-reported and measured sleep duration: how similar are they? *Epidemiology*. 2008;19(6):838–845. doi:10.1097/EDE.0b013e318187a7b0
7. Ibáñez V, Silva J, Cauli O. A survey on sleep assessment methods. *PeerJ*. 2018;6:e4849. doi:10.7717/peerj.4849
8. Quante M, Kaplan ER, Cailler M, et al. Actigraphy-based sleep estimation in adolescents and adults: a comparison with polysomnography using two scoring algorithms. *Nat Sci Sleep*. 2018;10:13–20. doi:10.2147/NSS.S151085
9. Atm VDW, Holmes A, Hurley DA. Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography - a systematic review: a systematic review of objective sleep measures. *J Sleep Res*. 2011;20(1pt2):183–200. doi:10.1111/j.1365-2869.2009.00814.x
10. De Zambotti M, Cellini N, Goldstone A, Colrain IM, Baker FC. Wearable sleep technology in clinical and research settings. *Med Sci Sports Exercise*. 2019;51(7):1538–1557. doi:10.1249/MSS.0000000000001947
11. Guillodo E, Lemey C, Simonnet M, et al. Clinical applications of mobile health wearable–based sleep monitoring: systematic review. *JMIR mHealth uHealth*. 2020;8(4):e10733. doi:10.2196/10733
12. Griffiths C, Silva KD, Hina F, et al. Effectiveness of a fitbit based sleep and physical activity intervention in an Early Intervention Psychosis (EIP) Service. *OJPsych*. 2022;12(02):188–202. doi:10.4236/ojpsych.2022.122015

13. Pengo MF, Won CH, Bourjeily G. Sleep in women across the life span. *Chest*. 2018;154(1):196–206. doi:10.1016/j.chest.2018.04.005
14. Maher C, Ryan J, Ambrosi C, Edney S. Users' experiences of wearable activity trackers: a cross-sectional study. *BMC Public Health*. 2017;17(1):880. doi:10.1186/s12889-017-4888-1
15. Peake JM, Kerr G, Sullivan JP. A critical review of consumer wearables, mobile applications, and equipment for providing biofeedback, monitoring stress, and sleep in physically active populations. *Front Physiol*. 2018;9:743. doi:10.3389/fphys.2018.00743
16. Zheng NS, Annis J, Master H, et al. Sleep patterns and risk of chronic disease as measured by long-term monitoring with commercial wearable devices in the All of Us Research Program. *Nat Med*. 2024;30(9):2648–2656. doi:10.1038/s41591-024-03155-8
17. Depner CM, Cheng PC, Devine JK, et al. Wearable technologies for developing sleep and circadian biomarkers: a summary of workshop discussions. *Sleep*. 2020;43(2). doi:10.1093/sleep/zsz254
18. de Zambotti M, Goldstone A, Claudatos S, Colrain IM, Baker FC. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiol Int*. 2018;35(4):465–476. doi:10.1080/07420528.2017.1413578
19. Doheny EP, Renerts K, Braun A, et al. Assessment of Fitbit Charge 4 for sleep stage and heart rate monitoring against polysomnography and during home monitoring in Huntington's disease. *J Clin Sleep Med*. 2024;20:1163–1171. doi:10.5664/jcsm.11098
20. Dong X, Yang S, Guo Y, Lv P, Wang M, Li Y. Validation of Fitbit Charge 4 for assessing sleep in Chinese patients with chronic insomnia: a comparison against polysomnography and actigraphy. *PLoS One*. 2022;17(10):e0275287. doi:10.1371/journal.pone.0275287
21. Eylon G, Tikotzky L, Dinstein I. Performance evaluation of Fitbit Charge 3 and actigraphy vs. polysomnography: sensitivity, specificity, and reliability across participants and nights. *Sleep Health*. 2023;9(4):407–416. doi:10.1016/j.sleh.2023.04.001
22. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Performance assessment of new-generation Fitbit technology in deriving sleep parameters and stages. *Chronobiol Int*. 2020;37(1):47–59. doi:10.1080/07420528.2019.1682006
23. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband fitbit models in assessing sleep: systematic review and meta-analysis. *J Med Internet Res*. 2019;21(11):e16273. doi:10.2196/16273
24. Onnela JP, Dixon C, Griffin K, et al. Beiwe: a data collection platform for high-throughput digital phenotyping. *JOSS*. 2021;6(68):3417. doi:10.21105/joss.03417
25. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry Res*. 1989;28(2):193–213. doi:10.1016/0165-1781(89)90047-4
26. Natale V, Esposito MJ, Martoni M, Fabbri M. Validity of the reduced version of the morningness-eveningness questionnaire. *Sleep Biol Rhythms*. 2006;4(1):72–74. doi:10.1111/j.1479-8425.2006.00192.x
27. Yu L, Buysse DJ, Germain A, et al. Development of short forms from the PROMIS sleep disturbance and sleep-related impairment item banks. *Behav Sleep Med*. 2011;10(1):6–24. doi:10.1080/15402002.2012.636266
28. Cole RJ, Kripke DF, Gruen W, Mullaney DJ, Gillin JC. Automatic sleep/wake identification from wrist activity. *Sleep*. 1992;15(5):461–469. doi:10.1093/sleep/15.5.461
29. Charge 3 101 Guide. Available from: https://device101.fitbit.com/guides/charge3-101.html. Accessed July 19, 2025.
30. Charge 5 101 Guide. Available from: https://device101.fitbit.com/guides/morgan-101.html. Accessed July 19, 2025.
31. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;327(8476):307–310. doi:10.1016/S0140-6736(86)90837-8
32. Godino JG, Wing D, De Zambotti M, et al. Performance of a commercial multi-sensor wearable (Fitbit Charge HR) in measuring physical activity and sleep in healthy children. *PLoS One*. 2020;15(9):e0237719. doi:10.1371/journal.pone.0237719
33. Menghini L, Yuksel D, Goldstone A, Baker FC, de Zambotti M. Performance of Fitbit Charge 3 against polysomnography in measuring sleep in adolescent boys and girls. *Chronobiol Int*. 2021;38(7):1010–1022. doi:10.1080/07420528.2021.1903481
34. McMahon M, McConley I, Hashim C, Schnyer DM. Fitbit validation for rest-activity rhythm assessment in young and older adults. *Smart Health*. 2023;29:100418. doi:10.1016/j.smhl.2023.100418
35. Degroote L, Hamerlinck G, Poels K, et al. Low-cost consumer-based trackers to measure physical activity and sleep duration among adults in free-living conditions: validation study. *JMIR mHealth uHealth*. 2020;8(5):e16674. doi:10.2196/16674
36. Ferguson T, Rowlands AV, Olds T, Maher C. The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study. *Int J Behav Nutr Phys Act*. 2015;12(1):42. doi:10.1186/s12966-015-0201-9
37. Schyvens AM, Van Oost NC, Aerts JM, et al. Accuracy of Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP Versus Polysomnography: systematic Review. *JMIR mHealth uHealth*. 2024;12:e52192–e52192. doi:10.2196/52192
38. Jaiswal SJ, Pawelek JB, Warshawsky S, et al. Using new technologies and wearables for characterizing sleep in population-based studies. *Curr Sleep Med Rep*. 2024;10(1):82–92. doi:10.1007/s40675-023-00272-7
39. Grandner MA, Rosenberger ME. Actigraphic sleep tracking and wearables: historical context, scientific applications and guidelines, limitations, and considerations for commercial sleep devices. *Sleep and Health Elsevier*. 2019:147–157. doi:10.1016/B978-0-12-815373-4.00012-5