Peer-reviewed version available at *Sensors* **2020**, *20*, 1378; doi:10.3390/s20051378

Article

Sleep in the natural environment: a pilot study

Fayzan F Chaudhry[1,2+], Matteo Danieletto[1,2,3+], Eddye Golden[1,2,3], Jerome Scelza[2,3], Greg Botwin[2,3], Mark Shervey[2,3], Jessica K. De Freitas[1,2,3], Ishan Paranjpe[1], Girish N. Nadkarni[1,4,5], Riccardo Miotto[1,2,3], Patricia Glowe[1,2,3], Greg Stock[3], Bethany Percha[2,3], Noah Zimmerman[2,3], Joel T. Dudley[2,3*], Benjamin S. Glicksberg[1,2,3*]

1 Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai, New York, NY, 10032, USA.
2 Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, 10032, USA
3 Institute for Next Generation Healthcare, Icahn School of Medicine at Mount Sinai, New York, NY, 10032, USA.
4 The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York.
5 Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, New York.

+ Authors contributed equally
* Correspondence: joel.dudley@mssm.edu or benjamin.glicksberg@mssm.edu

Abstract: Sleep quality has been directly linked to cognitive function, quality of life, and a variety of serious diseases across many clinical domains such as psychiatry and cardiology. Standard methods for assessing sleep involve overnight studies in hospital settings, which are uncomfortable, expensive, not representative of real sleep, and difficult to conduct on a large scale. Recently, a number of commercial digital devices have been developed that record physiological data which can act as a proxy for sleep quality in lieu of standard electroencephalogram recording equipment. Each device company makes different claims of accuracy and measures different features of sleep quality, and it is still unknown how well these devices correlate with one another and perform in a research setting. In this pilot study of 21 participants, we investigated whether outputs from four sensors, specifically FitBit, Withings Aura, Hexoskin, and Oura Ring, were related to known cognitive and psychological metrics, including the PSQI and N-back test. We found that sleep metrics extracted from these devices did not predict cognitive and psychological metrics well in our pilot data. However, we did identify certain signification associations, specifically the Oura Ring's total sleep duration and efficiency in relation to the PSQI measure with p=0.004 and p=0.033, respectively. Additionally, correlation of various sleep features among the devices across the sleep cycle was almost uniformly low. These findings can hopefully be used to guide future sensor-based sleep research.

Keywords: Wearables, Biosensors, Sleep, Fitbit, Oura, Hexoskin, Withings, Cognition

1. Introduction

Between 50 and 70 million Americans currently suffer from poor sleep [1]. A 2014 study from the Centers for Disease Control and Prevention (CDC) found that over one third of Americans (34.8%) regularly sleep less than the recommended 7 hours per night [2]. Although the body has remarkable compensatory mechanisms for acute sleep deprivation, chronic poor sleep quality and suboptimal sleep duration are linked to many adverse health outcomes, including increased risk of diabetes [3], metabolic abnormalities [4], cardiovascular disease [5], hypertension [6], obesity [7], anxiety [8], and depression [8]. Chronic sleep deprivation also burdens society economically, contributing to premature mortality, loss of working time, and suboptimal education outcomes that cost the US $280.6-411 billion annually [9]. However, the underlying mechanisms mediating the adverse effects of poor sleep remain unknown. Diverse factors and complex interactions govern the relationship between health and sleep, and there is likely substantial inter-individual variability. Pronounced gender [10], race [11], and ethnicity differences in sleep-related behaviors are well-established [2].

It is clear that broad, population-level studies of sleep are necessary to understand how lifestyle and environmental factors contribute to poor sleep and to link sleep abnormalities to their attendant negative health effects [12]. It is particularly important to capture individuals' sleep patterns in natural sleep settings: i.e. at home. However, traditional approaches to studying sleep do not permit these types of studies. Polysomnography (PSG), where brain waves, oxygen levels, eye and leg movements are recorded, is the current "gold standard" approach to studying sleep. A PSG study typically requires the participant to sleep in a hospital or clinic setting with uncomfortable sensors placed on the scalp, face, and legs. These studies, which remove the participant from his/her natural sleep environment, are not well suited to longitudinal assessments of sleep. They also create issues such as First Night Effects (FNE), which limit the translatability of laboratory sleep studies to real-life environments [13]. The recent development of clinical grade, at-home PSG tools has enabled quantification of the laboratory environment's effect on sleep [14]. These studies have generally confirmed that participants sleep better at home than they do in a lab [15, 17], although these findings are not universal [15].

Even with the availability of at-home PSG, it is unlikely that the use of expensive, cumbersome, single-purpose equipment will promote the kinds of large-scale population studies that can quantify the diverse factors affecting sleep and its relationship to health outcomes. More user-friendly, lightweight, and unobtrusive sleep sensors are needed; ideally these will be embedded in devices that study participants already own. Recently, several companies have developed sub-clinical grade "wearable" technologies for the consumer market that passively collect high frequency data on physiological, environmental, activity, and sleep variables [16]. The FDA classifies these as general wellness products and they are not approved for clinical sleep studies. Due to their passivity, low risk, and growing ubiquity amongst consumers, however, it is clear that these devices present an intriguing new avenue for sleep data collection at scale [17]. Combined with mobile app software to monitor cognitive outcomes such as reaction time, executive function, and working memory, these devices could feasibly be used for large-scale, fully remote sleep studies.

Here we present results from a week-long pilot comparison study of four commercially available wearable technologies for sleep monitoring. Twenty-one participants were instrumented with all four

devices for the entire week. To assess the feasibility of a fully remote study relating sleep features to cognition, we also assessed participants' cognitive function daily via a series of assessments on a custom-built mobile app. None of the four devices we compared in this study have been compared head-to-head before for sleep and cognition research. Our results, which show low correlation among devices as well as low correlation between quantified sleep metrics and cognitive outcomes, highlight some of the key difficulties involved in designing and executing large-scale sleep studies with consumer-grade wearable devices.

2. Materials and Methods

2.1. Research setting

Participants were enrolled individually at the Harris Center for Precision Wellness (HC) and Institute for Next Generation Healthcare research offices within the Icahn School of Medicine at Mount Sinai. Monetary compensation in the form of a $100 gift card was provided to study participants upon device return. During the enrollment visit, participants met with an authorized study team member in a private office to complete the consent process, onboarding, and baseline procedures. The remainder of the study activities took place remotely with limited participant-team interaction. The study team maintained remote contact with each research participant throughout his/her participation via phone or email to answer any questions and provide technical support. The study was approved by the Mount Sinai Program for the Protection of Human Subjects (IRB #15-01012).

2.2. Recruitment methods

We employed a variety of recruitment methods to ensure a diverse sample population.

1. IRB-approved posting on the Mount Sinai trials website (http://icahn.mssm.edu/research/clinical-trials)
2. Advertisements about our work placed in local Mount Sinai media and flyers (e.g., department monthly newsletters, flyers hung in appropriate areas in Mount Sinai, Mount Sinai Weekly Academic Update, etc.)
3. IRB-approved advertisement on Internet websites (e.g. Dudley Lab http://dudleylab.org; Harris Center http://www.precisionwellness.org; Icahn Institute http://icahn.mssm.edu/departments-and-institutes/genomics) and other websites that help match studies with participants in clinical trials [e.g., Research Match (https://www.researchmatch.org), and CenterWatch).
4. Word-of-mouth from individuals previously enrolled in other Harris Center trials
5. Social Medial postings (e.g. Facebook, Twitter)
6. Facebook ads, Google Adwords, and other advertising outlets online, craigslist.org and other community bulletin board systems, patientslikeme.org and other community and support group websites and networking sites.

2.3. Inclusion and exclusion criteria

Participants were eligible for the study if they were over 18 years old, had access to an iPhone, had basic knowledge of installing and using mobile applications and wearable devices, and were willing and able to provide written informed consent and participate in study procedures.

Participants were ineligible for the study if they were colorblind, due to initial strop test restrictions that were not utilized, part of a vulnerable population, or unwilling to consent and participate in study activities.

2.4. Onboarding questionnaires

During the initial study visit, participants were prompted to take four questionnaires (see Appendices A-D). All questionnaires were completed electronically via SurveyMonkey and the results were subsequently stored in the study team's encrypted and secured electronic database.

The Demographics Questionnaire (Appendix A) ascertained basic demographic information.

The 36-Item Short Form Health Survey (SF-36; Appendix B). The SF-36 evaluates eight domains: physical functioning, role limitations due to physical health, role limitations due to emotional problems, energy/fatigue, emotional well-being, social functioning, pain and general health. The SF-36 takes roughly 5-10 minutes to complete.

The Morningness-Eveningness Questionnaire (MEQ; Appendix C) is a 19-question multiple-choice instrument designed to detect when a person's circadian rhythm allows for peak alertness. The MEQ takes roughly 5-10 minutes to complete.

The Pittsburgh Sleep Quality Instrument (PSQI; Appendix D) is a 9-item, self-rated questionnaire that assesses sleep over the prior month. The PSQI has been shown to be sensitive and specific in distinguishing between good and poor sleepers. The PSQI utilizes higher numbers to indicate poorer sleep. The PSQI takes roughly 5-10 minutes to complete.

2.5. Technology set-up and testing

After the initial screening visit, participants were asked to set up their devices and begin the week-long study at their leisure (Figure 1). The study team chose technologies based on performance and usability data obtained from HS#: 15-00292, "Pilot Evaluation Study on Emerging Wearable Technologies." Each participant was assigned four sleep monitoring devices: a first edition Fitbit Surge smart watch, a Hexoskin smart shirt, a Withings Aura sleep pad/system, and a first edition Oura smart ring. Note that the form factors for the four devices were different; this was important to ensure that they could all be used at once and would not interfere with each other.

Setup for each device involved downloading the corresponding manufacturer's mobile application on the participant's iPhone and downloading the study team's custom "HC App". Participants agreed to each manufacturer's software terms and conditions in the same manner as if they were to purchase and install the technologies themselves. In doing so, and as noted in the participant-signed consent document, participants acknowledged that the manufacturers would have access to identifiable information such as their names, email addresses, and locations.

The HC App functioned as a portal to allow participants to authorize the sharing of data between the manufacturers' applications and the study team's database. During the initial set-up period, the

study team worked with participants to troubleshoot any issues and ensure proper data transmission to the database.

2.6. Sleep monitoring and device-specific parameters

Over a 7-day consecutive monitoring period of the participant's choosing, participants used the four different sleep monitoring technologies and completed daily assessments (Figure 1). The monitors measured physiological parameters (e.g. heart rate, heart rate variability, respiratory rate, temperature, and movement), activity parameters (e.g. number of steps per day), and sleep related parameters (e.g. time in each sleep stage, time in bed, and number of wake ups per night). Participants were also given the option of wearing two monitors (Fitbit Surge, Oura smart ring) during the day over this 7-day period.

The Withings and Oura both stage sleep as: (1) awake, (2) light, (3) deep, and (4) REM (Figure 1). The Hexoskin stages sleep as (1) awake, (2) NREM, or non-REM, and (3) REM. The Fitbit stages sleep as (1) very awake, (2) awake, and (3) asleep.

2.7. Daily questionnaires and N-back tests

Using the HC App, participants completed questionnaires and cognitive assessments on each day of the 7-day study.

N-back tests: The N-back test [18] assesses working memory as well as higher cognitive functions/fluid intelligence. Participants were prompted to take the N-back test three times per day (morning, afternoon, and evening). In each test, participants were presented with a sequence of 20 trials, each of which consisted of a picture of one of eight stimuli: eye, bug, tree, car, bell, star, bed, anchor. The participant was asked whether the image was the same as the image "N"-back from the current image, where N = 1 or 2. The stimuli were chosen so that in the course of 20 trials, 10 would be a "hit" (the stimulus would match the N-back stimulus) and 10 would be a "miss". The participant had 500 ms to enter a response. If no response was entered, the trial was counted as incorrect and a new trial was presented. The N-back tests took roughly 3 minutes each, for a total of under 10 min/day.

Modified PSQI for Daily Use (MD-PSQI): We modified the PSQI so that it could be administered daily via the HC App, enabling a low-burden user assessment of his/her own sleep quality the previous night. the MD-PSQI asked the user for an estimate of total sleep duration (TSD; i.e., total amount of sleep), latency (i.e., time to fall asleep), and start to end sleep duration (i.e., TSD plus latency). Participants took the MD-PSQI electronically through the HC App at wakeup (1-2 minutes completion time).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2020 doi:10.20944/preprints202002.0046.v1

Peer-reviewed version available at Sensors 2020, 20, 1378; doi:10.3390/s20051378

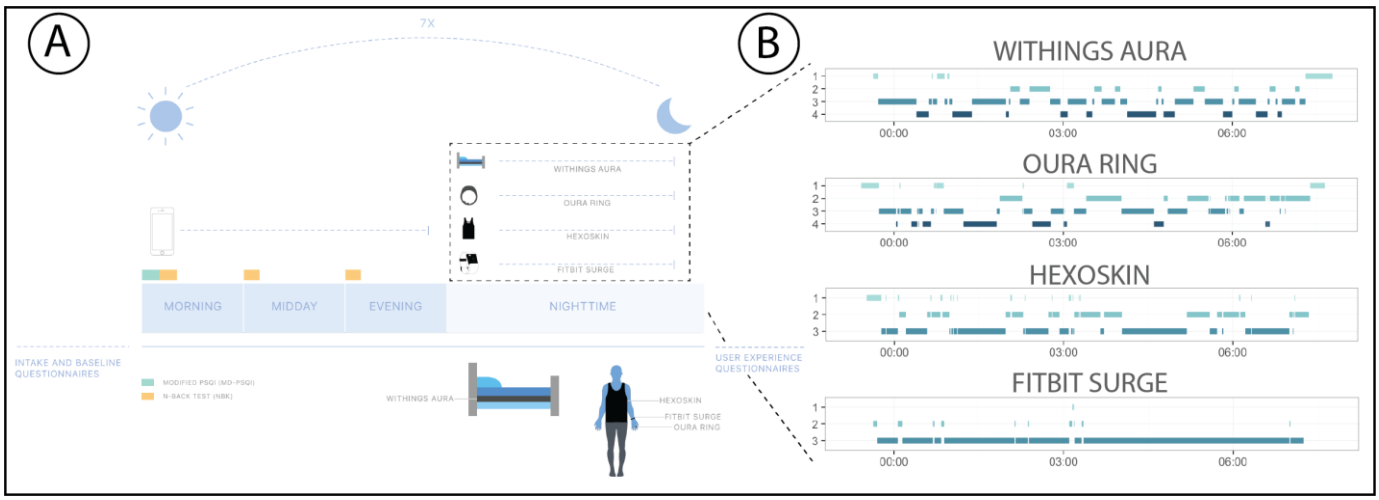

Figure 1. Study structure and data collection for our pilot sleep study. A: Illustration of sleep study monitoring procedure and data collection strategy. B: Example data showing a comparison of sleep staging of a single night for one study participant for all four devices.

2.8. N-back test scoring

For each trial, the participant's response time and the correctness/incorrectness of responses were recorded. We calculated four different scores for the N-back tests: median reaction time and percent correct, stratified by congruent vs. incongruent items. We treated all reaction times the same and did not segment or weight based on items that the participant got correct vs. incorrect. Only morning test results were used in the final analysis, as we theorized that sleep-related factors would affect morning results most strongly. Each participant was then given a cognition score based on a self-created scoring function (Eq. 1) of the reaction time, degree of difficulty of question, and correctness. The metric accounts for variation across multiple elements of the N-back results leading to a greater representation of performance. The formula for the metric is

$$\frac{\sum\left(1-\frac{Reaction\ Time}{Max\ Reaction\ Time}\right)*Answer\ Correct*\frac{Steps\ Back}{2}}{N} \quad (1)$$

2.9. Device comparisons

We compared the four devices on a per-epoch basis, calculating the Pearson correlation coefficient between each pair of devices per participant-night. Pearson was chosen after analysis of variable density plots and utilization of a more robust outlier centric correlation metric was deemed unnecessary. Because we were interested only in concordance between devices, our calculations did not take into account groups of nights corresponding to the same participant. We recorded the median, minimum, and maximum correlations between each pair of devices over all participant-nights. We also compared the sleep metrics produced by the devices to each participant's self-assessment of his/her total sleep duration, sleep latency (how long it took to fall asleep) and sleep efficiency (i.e., TSD divided by total time in bed). Three of the four study devices calculated these metrics automatically, and we manually calculated them for the fourth device (Withings).

2.10. Models linking device data to MD-PSQI and N-back scores

We built a series of univariate linear models that regressed each individual sleep feature on either (a) PSQI score or (b) N-back score. The PSQI tracks quality of sleep with higher values indicating poorer sleep. We regressed the one-time reported PSQI on all available device metrics, taking the mean of each metric across all nights of sleep for each individual. Additionally, we regressed the N-Back created score to show variance in cognition score across all mean device data by subject and self-reported data. In all of the regression models for both PSQI score and N-back score, we analyzed only the data from individuals with two or more days of reported N-back scores, leaving us with 16 of the original 21 individuals.

2.11. Analysis of missing data

We analyzed the degree of missingness of each (a) device-reported or (b) user-reported field as measures of (a) device reliability/quality or (b) participant compliance. As the study progressed, some sleep features were also updated due to new advances in hardware and software on the device side, which resulted in missing data columns, which were not included in the missing data plot.

3. Results

3.1. Summary of study population

Table 1 describes our study population, which consisted of 21 participants (11 female; 10 male). The median age of our population was 29 (range: 23 to 41). The median PSQI score was 4 (range: 1 to 12). Sixteen of our participants were classified as normal sleepers, three were poor sleepers, and two were very poor sleepers. Median MEQ score was 52 (range: 35 to 73). We provide score summaries for all eight SF-36 subcategories at the bottom of Table 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2020 doi:10.20944/preprints202002.0046.v1

Peer-reviewed version available at Sensors 2020, 20, 1378; doi:10.3390/s20051378

Table 1. Summary of the study population. We include the participant's gender (M/F/O), baseline assessment of sleep quality according to the PSQI with higher values indicative of poorer sleep, age, SF-36 score (a measure of general health along 8 axes), and MEQ time (optimal time of day). SF-36 items with a perfect score of 100 were replaced by "-" to improve table readability.

ID	Gender	Age	PSQI	MEQ	SF-36 Scores							
					Physical Functioning	Role Limitations (Physical)	Role Limitations (Emotional)	Energy	Emotional Well-being	Social Functioning	Pain	General Health
1	F	23	1	50	-	-	-	50	68	87.5	-	55
2	F	26	4	47	90	-	66.7	45	72	-	-	60
3	F	27	5	52	-	-	-	45	56	87.5	90	50
4	F	27	2	36	-	-	-	65	80	75	-	55
5	F	27	4	58	-	-	-	50	76	87.5	90	55
6	F	28	3	52	-	-	-	55	76	75	-	60
7	F	28	3	40	90	-	33.3	50	72	87.5	67.5	55
8	F	29	12	35	-	-	0	15	36	50	67.5	55
9	F	31	4	49	95	-	-	60	84	-	-	55
10	F	39	4	49	60	50	-	45	44	87.5	77.5	55
11	F	41	5	53	-	-	-	95	96	-	-	60
12	M	25	10	55	-	-	66.7	85	76	-	-	60
13	M	29	5	52	-	-	-	50	88	-	-	50
14	M	29	4	41	-	-	-	50	76	-	-	60
15	M	31	3	56	95	-	-	65	80	75	90	50
16	M	34	12	73	-	-	66.7	50	52	62.5	-	55
17	M	35	6	52	-	-	-	75	80	-	-	50
18	M	37	3	61	90	-	66.7	50	80	87.5	90	55
19	M	39	8	72	-	-	-	80	88	-	-	55
20	M	41	6	55	95	-	-	50	84	-	80	55
21	M	41	9	52	95	-	66.7	35	52	87.5	70	60
MIN		23	1	35	60	50	0	15	36	50	67.5	50
MEDIAN		29	4	52	100	100	100	50	76	87.5	100	55
MAX		41	12	73	100	100	100	95	96	100	100	60

3.2. Device comparisons

Total sleep duration (TSD) was reported by three of the devices and by the participants themselves. Figure 2a shows a correlation matrix of total sleep duration. The correlations were generally medium to weak ($\rho < 0.7$ for all pairwise comparisons), although surprisingly the correlations of the user-generated estimates with device estimates were on par with correlations among the devices themselves. Figure 2b shows a REM sleep (in seconds) cycle correlation across the Oura, Hexoskin and Withings (Fitbit did not report an estimate of REM sleep). The correlation between Oura and Withings was highest at $\rho = 0.44$, while Oura and Hexoskin were least correlated ($\rho = 0.22$). Figure 2c

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2020 doi:10.20944/preprints202002.0046.v1

Peer-reviewed version available at Sensors 2020, 20, 1378; doi:10.3390/s20051378

shows a sleep cycle tracking correlation matrix. The most highly correlated devices were the Hexoskin and Oura at ρ = 0.4. The least correlated devices were Withings and Fitbit at ρ = 0.12.

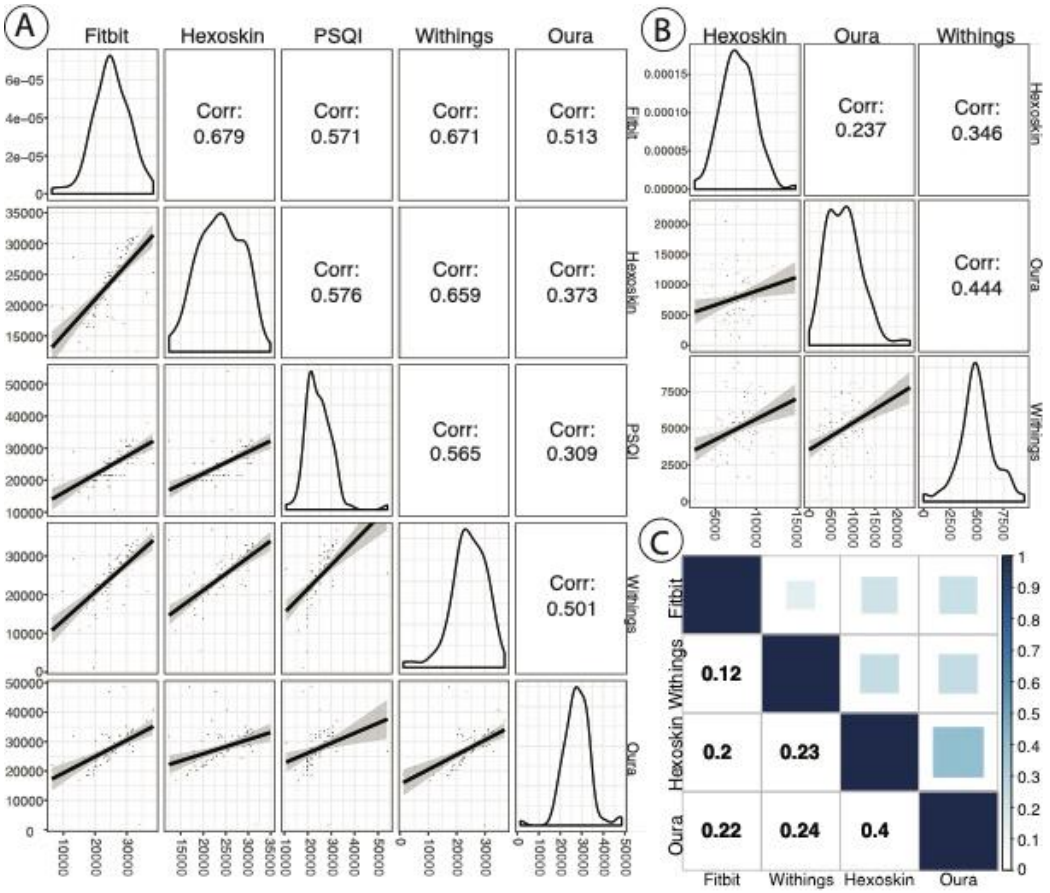

Figure 2. A: A correlation matrix of total sleep duration (in seconds) by device. Each point represents one participant-night. B: A REM sleep (in seconds) cycle correlation across the Oura, Hexoskin and Withings. The Fitbit was excluded, as it does not track REM vs. non-REM sleep. C: A sleep cycle tracking correlation matrix. Oura and Withings track four sleep stages while the Hexoskin and Fitbit track three stages. The Fitbit has one stage for sleep and two stages for non-sleep.

Table 2. Summary metrics of device data. All units are in hours except wakeups which is in occurrences and efficiency which has no units. Sleep efficiency is a metric to track percentage of time in bed while asleep. TSD is total sleep duration which is similar to start-end duration and variants were utilized that included latency and other measures.

Device	Metric	N	Mean	St. Dev	Min	Pctl (25)	Pctl (75)	Max
Fitbit	Efficiency	129	94.70	15.70	31.00	94.00	97.00	193.00
	TSD All	129	7.47	1.47	3.78	6.50	8.43	11.40
	TSD	129	7.58	1.58	1.78	5.98	7.93	10.75
	Start-End	129	7.58	1.73	3.78	6.50	8.48	15.87
	Wakeups	129	1.60	1.20	0.00	1.00	2.00	8.00
Hexoskin	Efficiency	114	92.40	4.40	70.30	91.10	95.30	97.80

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2020 doi:10.20944/preprints202002.0046.v1

Peer-reviewed version available at Sensors 2020, 20, 1378; doi:10.3390/s20051378

	TSD	114	6.72	1.31	3.45	5.78	7.81	9.69
	Start-End	135	7.57	1.42	3.93	6.57	8.58	11.43
	REM	123	2.15	0.57	0.69	1.77	2.53	4.12
	Latency	114	0.29	0.26	0.07	0.12	0.38	1.56
Oura	Efficiency	127	89.70	14.40	24.00	84.00	93.00	164.00
	TSD	128	7.69	1.72	0.42	6.73	8.75	13.48
	Start-End	130	10.67	11.63	4.62	6.97	9.55	117.60
	REM	127	2.17	1.11	0.00	1.29	2.81	6.38
	Deep	127	1.12	0.58	0.00	0.73	1.44	2.58
	Wakeups	127	2.40	1.90	0.00	1.00	4.00	7.00
	Latency	127	0.26	0.25	0.01	0.11	0.30	1.58
Withings	Efficiency	141	84.10	20.50	20.50	74.80	90.10	179.80
	TSD All	141	8.99	2.89	0.53	7.45	10.12	27.03
	TSD	141	6.97	1.75	0.33	5.95	8.15	10.97
	Start-End	141	9.30	4.45	0.42	7.08	9.73	34.55
	REM	141	1.40	0.46	0.00	1.15	1.67	2.63
	Deep	141	1.74	0.58	0.00	1.42	2.15	3.67
	Light	141	3.83	0.98	0.33	3.22	4.45	6.03
	Wakeups	141	2.40	2.60	0.00	0.00	3.00	13.00
	Latency	141	0.32	0.36	0.00	0.08	0.42	2.37
	Wakeup Duration	141	1.38	2.14	0.03	0.53	1.50	17.48
MD-PSQI	Start-End	122	7.34	1.45	4.50	6.35	8.24	12.33
	TSD	122	6.91	1.56	3.00	6.00	7.78	15.00
	Latency	122	0.24	0.23	0.02	0.08	0.33	2.00

3.3. PSQI and cognition score vs. device data

Table 3 shows the results of several univariate linear models, each of which included either PSQI or cognition score as the dependent variable and a different mean device measure per subject as the independent variable. The only statistically significant associations (at a significance threshold of alpha = 0.05) were between Oura's measurement of total sleep duration and sleep efficiency and PSQI ($p < .01$ for both). In both cases, an increase in total sleep duration or sleep efficiency was associated with a significant decrease in PSQI score; since PSQI increases with poor sleep quality, these associations are in the expected direction (more sleep or more efficient sleep leads to better or lower PSQI). There were no significant associations between cognition score and any of the device sleep variables.

3.4. Cognition score vs. participant summary data

Table 4 shows the results of univariate linear models that regressed cognitive score on participant summary features. The SF-36 sub-category "physical functioning" had a significant association with

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2020 doi:10.20944/preprints202002.0046.v1

Peer-reviewed version available at Sensors 2020, 20, 1378; doi:10.3390/s20051378

cognition score (p = 0.014); however, further analysis revealed that this was due to the presence of an outlier with very low physical functioning as well as a low cognition score, and removal of this individual destroyed the effect. "Emotional well-being" had a weakly significant association (p = 0.078) with cognition score that appears robust to the removal of individual data points. None of the other summary features were significantly associated with cognition score.

Table 3. Results of multiple univariate Linear Models. The leftmost univariate linear models are the mean device data by subject as independent variables and PSQI as the dependent variable. Please note the higher the value is on the PSQI the worse the sleep quality, thus positive correlations suggest relation to poorer sleep quality. The rightmost model is the mean device data by subject as independent variables and Cognition Score as the dependent variable. All units are in hours with the exception of Wakeups (occurrences) and Efficiency (a standardized metric).

		PSQI				Cognition Score			
Device	Feature	Coefficient	Std. Error	P-Value	R^2	Coefficient	Std. Error	P-Value	R^2
Fitbit	TSD	-0.273	.544	.622	.014	-.003	.014	.825	.004
	Wakeups	1.570	1.005	.136	.119	-.018	.018	.329	.068
Withings	TSD	-0.125	.498	.804	.004	-.018	.010	.110	.172
	Latency	-2.08	2.83	.472	.0291	.015	.091	.869	.002
	Efficiency	0.010	.060	.869	.002	-.001	.001	.315	.072
	Wakeups	.352	.427	.421	.036	.001	.001	.888	.001
	REM	.260	1.962	.896	.001	-.039	.040	.342	.065
Oura	TSD	-1.004	.305	.004***	.376	-.022	.019	.265	.088
	Latency	-7.311	4.445	.117	.131	-.105	.113	.366	.059
	Efficiency	-.092	.0398	.033**	.228	.002	.002	.285	.081
	Wakeups	.168	.491	.736	.006	-.018	.014	.226	.103
	REM	-0.526	0.715	.471	.029	-.008	.0173	.656	.015
Hexoskin	TSD	.187	.702	.793	.004	-.020	.015	.206	.112
	Latency	1.249	4.444	.782	.004	-.001	.096	.995	.000
	Efficiency	-0.226	.272	.417	.037	.003	.005	.530	.029
	REM	.397	1.833	.831	.003	-.071	.044	.128	.158
MD-PSQI	TSD	-0.725	.558	.210	.086	-.005	.014	.725	.009
	Latency	1.846	4.033	.653	.012	-.008	.091	.935	.000
Observations		16				16			

Note: *p<0.1; **p<0.05; ***p<0.01

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2020 doi:10.20944/preprints202002.0046.v1

Peer-reviewed version available at Sensors 2020, 20, 1378; doi:10.3390/s20051378

Table 4. In this collection of univariate linear models, the participant summary data are independent variables and Cognition Score is the dependent variable. These metrics represent all scores that have been standardized.

	Cognition Score			
Feature	Coefficient	Std. Error	P-Value	R^2
PSQI	-.003	.005	.531	.029
MEQ	-.001	.002	.529	.029
Emotional Role Limitations	.0003	.001	.665	.014
Energy	.0003	.001	.700	.011
General Health	-.001	.004	.769	.006
Physical	.003	.001	.014**	.358
Social	.002	.001	.170	.130
Well-being	.002	.001	.078*	.206
Observations	16			

Note: *p<0.1; **p<0.05; ***p<0.01

3.5. Correlation of MEQ preference with cognitive test response rates

We illustrate the rate of missingness for sleep-related metrics in Figure 3; in general, a significant proportion of relevant data are missing due to noncompliance by users or device malfunctioning. We stratified response rate for morning, afternoon and evening test results and displayed these for participants grouped by their MEQ segmentation into Night, Intermediate, and Morning in Figure 4. We see that morning-preferred participants had the lowest response rate across all times. Furthermore, we see that afternoon response times are the highest for all MEQ groupings.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 February 2020 doi:10.20944/preprints202002.0046.v1

Peer-reviewed version available at Sensors 2020, 20, 1378; doi:10.3390/s20051378

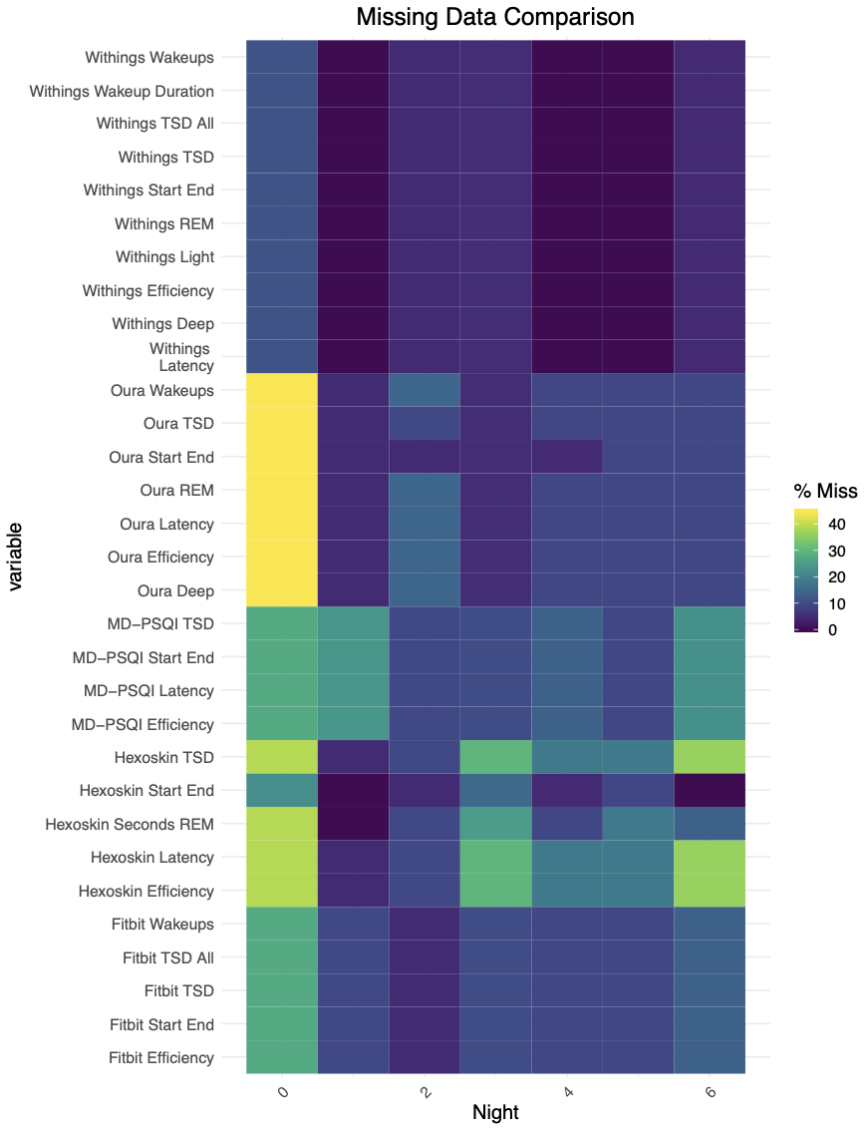

Figure 3. Plot of missing sleep-related data including the modified PSQI (MD-PSQI). Due to various device preferences missing data is asymmetric across devices.

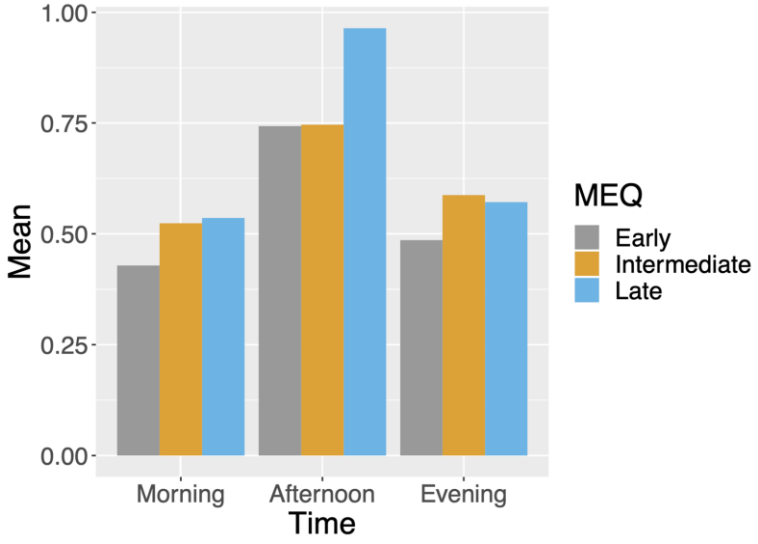

Figure 4. Average missing data for N-back test by time (Morning, Afternoon, and Evening) and MEQ groupings (Early, Intermediate, or Late).

4. Discussion

The results of our study reflect some general findings that are likely to impact most research involving wearable devices and mobile apps. First, because of low enrollment, our ability to detect effects was low; an effect would need to be highly pronounced to be detectable in a study population of this size. The effort involved in publicizing the study, enrolling participants, and ensuring they were able to complete the study (no device or app malfunctions, devices running out of batteries, etc.) was substantial. Simple study designs with perhaps 1-2 devices that participants already own and are familiar with offer the greatest chance of success on a large scale. Second, there was substantial variability among the devices we tested, making the choice of device for any sleep study a material factor that can impact results. Even if it's impossible to assess which device is "preferred" for a given study design, this variability impacts cross-interpretability of results across different studies and will thwart attempts at meta-analyses. Third, missingness and the presence of outliers were important considerations for all statistical analyses on this dataset. Although this was a pilot study, all of these issues are likely to translate to larger wearable device studies as well.

Related work and study limitations: sleep-monitoring devices

This pilot study consisted of 21 participants whose sleep was tracked for seven days across four devices. This built off of previous work that utilized comparisons of various devices and polysomnography[20,28]. In previous work, de Zambotti directly compared the Oura ring with PSG. Correlation matrices from their study show poor agreement across different sleep stages, showing that tracking sle

ep stages was a problem for the Oura. However, this study concluded the Oura's tracking of total sleep duration, sleep onset latency, and wake after sleep onset were not statistically different than that of PSG. In the current study, for total sleep duration we obtained correlations with the Oura of 0.513, 0.373, and 0.501 with the Fitbit, Hexoskin, and Withings, respectively. The Oura is a good indication for sleep duration and latency because it was found to track total sleep duration in relative accordance with the PSG. This suggests that many devices have trouble tracking total sleep duration or participants had trouble wearing devices correctly outside of a monitored sleep lab. Across the three devices that tracked REM, the maximum correlation was 0.444 between the Oura and the Withings. One noteworthy observation is that consensus correlation of sleep cycles between the Oura and the Hexoskin was 0.4 despite lack of standardization of sleep cycles.

The biggest question for these devices is, how well do they actually reflect sleep? The current consensus is mixed. For instance, de Zambotti et al. found good overall agreement between PSG and Jawbone UP device, but there were over- and underestimations for certain sleep parameters such as sleep onset latency [19]. Another study compared PSG to Oura ring and found no differences in sleep onset latency, total sleep time, and wake after sleep onset but they did find differences in sleep stage characterization between the two recording methods [20]. Meltzer et al. [21] concluded that the Fitbit Ultra did not produce clinically comparable results to PSG. Montgomery-Downs et al. [22] found that Fitbit and actigraph monitoring consistently misidentified sleep and wake states compared to PSG and they highlight the challenge of using such devices for sleep research in different age groups. While such wearables offer huge promise for sleep research, there are a wide variety of additional challenges that exist of their utility, including accuracy of sleep automation functions, detection

range, tracking reliability, among others [23]. Furthermore, comprehensive research including randomized control trials as well as interdisciplinary input from physicians and computer, behavioral, and data scientists will be required before these wearables can be ready for full clinical integration[24].

As there are many existing commercial devices, it is not only important to determine how accurate they are in capturing certain physiological parameters, but also the extent to which they are calibrated compared to one another. In this way, findings from studies that use different devices but measure similar outcomes can better be taken together in context. Murakami et al. [25] evaluated 12 devices in their ability to capture total energy expenditure against gold standard and found that while most devices had strong correlation (greater than 0.8) compared to the gold standard, they did vary in their accuracy with some significantly under- or overestimating energy expenditure. The authors suggest that most wearable devices do not produce a valid quantification of this endpoint. Xie et al. [26] compared six devices and two smartphone apps in their ability to measure major health indicators (e.g., heart rate, number of steps) under various activity states (e.g., resting, running, sleeping). They found that the devices had high measurement accuracy for all health indicators except energy consumption, but there was variation between devices and certain ones performed better than others for specific indicators in different activity states. In terms of sleep, they found the overall accuracy for devices to be high in comparison to output from Apple Watch 2, which was used as the gold standard. Lee et al. [27] performed a highly relevant study in which they examined the comparability of five devices total and a research-grade accelerometer to self-reported sleep in their ability to capture key sleep parameters such as total sleep time and time spend in bed for one to three nights of sleep.

The biggest limitation of the current study was the lack of a gold standard for sleep metrics, namely PSG. It should be noted that sleep studies are extremely hard to conduct on a high number of participants due to the prohibitive cost of PSG. In the future, however, this field can face huge growth if some amalgamation of cheap, at home devices could reliably track various data and cross confirm results amongst themselves. This would be extremely beneficial in creating a mapping function of individual device metrics to PSG metrics which in turn could allow these more simplistic sensors to accurately recreate conditions of PSG at a low cost and in the comfort of the home. This mapping function could increase recruitment of participants while decreasing cost for sleep studies.

Considerations related to cognitive metrics and self-reported sleep quality indices

Self-reported PSQI has been shown to be a poor screening measure of PSG[29]. This may help explain why self-reported one-time PSQI sleep quality variation was not well explained by much of the device data. However, the Oura ring efficiency and sleep duration did explain variation in the one-time self-reported PSQI with statistical significance. These Oura tracking metrics may merit further investigation. Also, it is important to note that poor tracking metrics, or a low number of participants could also be the reason more device data was not able to explain variation in PSQI. For the modified daily PSQI survey, having participants track their individual sleep metrics such as latency may not have been as useful as a simple overall estimate of quality of sleep.

Evidence of using the N-back test as a fluid intelligence metric is contentiously accepted, with some critics citing low correlation between N-back and other fluid intelligence tests[30]. The

cognition metric, taken from the N-back results, and results from participant summary data were not strongly correlated, most likely due to study methods or a low sample size. A recent study showed that poor sleep or deprivation may cause local deficits specifically with tasks emotional in nature[31]. This may suggest implementing a metric for morning or afternoon wellbeing in addition to fluid intelligence tasks.

Insight from response rate based on MEQ segmentation into three categories early, intermediate and late preference individual could help future study design. We found that response rates across the afternoon preference individuals was highest across all MEQ groups. This suggests that crucial surveys should be administered around this time if possible. Another finding of note is that evening preference participants had the best response rate across the morning/afternoon and were close to the highest in the evening. This finding suggests that non-evening preference participants may need extra motivating factors to increase their response rates. Considering the high rate of evolution for wearable device and new devices coming to market monthly data portability is an important future problem.

5. Conclusion

We reported correlations among important sleep metrics for four different sleep tracker devices and correlated the results with self-reporting questionnaires and cognitive metrics, specifically the N-back. Difficulty in participant enrollment and engagement led to new ideas about recruitment design and participant engagement design. Exploiting existing technology such as ReasearchKit, HealthKit from Apple can have a twofold benefit for recruiting people remotely with e-consent feature from ResearchKit and sharing electronic health records (EHR). Further combining this with additional data stores present in the HealthApp can improve the eligibility of participants[32]. In consideration of the missing data in the questionnaires and active tasks prescribed, we promote the use of as passive collection procedures as possible. One such option is a smart mirror [33], which is even more natural than using the smartphone because the mirror is already incorporated into the daily routine. Finally, the weak correlation among devices opens new challenges for accurate interpretation and data portability for the end user. How will device-specific findings from various studies be taken in context to one another? The results from the current study can hopefully highlight the need for better standardization for sleep-related metrics across devices to make any robust and accurate conclusions. These considerations are especially important in this era of digital health, where device firmware is continually updated and individuals frequently upgrade to new sensors.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Appendix A: Demographic questionnaire, Appendix B: SF-36 questionnaire, Appendix C: MEQ questionnaire, Appendix D: PSQI questionnaire.

Author Contributions: All authors have read and agree to the published version of the manuscript. Conceptualization, G.S., N.Z., B.P., and J.T.D.; methodology, M.S., J.S., B.P., and N.Z.; software, J.S., D.S., M.S., and G.S.; formal analysis, F.F.C, M.D., and B.S.G.; investigation, G.B. and E.G.; validation, R.M, M.D., J.K.D-F., I.P., and G.N.N.; data curation, F.F.C., M.D., J.S., M.S., and N.Z.; writing—original draft preparation, F.F.C., M.D., E.G., B.P., N.Z., and B.S.G; writing—review and editing, all authors.; visualization, F.F.C., M.D., J.S., N.Z.,

and B.S.G.; supervision, J.T.D. and B.S.G.; project administration, G.B., E.G., and P.G.; funding acquisition, G.S. and J.T.D.

Funding: This research, and The Harris Center for Precision Wellness, was funded by generous gifts from Joshua and Marjorie Harris of the Harris Family Charitable Foundation and Julian Salisbury.

Acknowledgments: We acknowledge David E. Stark for assistance in designing the cognitive battery and Christopher Cowan for design and development of the HC mobile app.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

5. References

1. Research, N.C.o.S.D. National Institutes of Health sleep disorders research plan. *Bethesda, MD: National Institutes of Health* **2011**.
2. Liu, Y.; Wheaton, A.G.; Chapman, D.P.; Cunningham, T.J.; Lu, H.; Croft, J.B. Prevalence of Healthy Sleep Duration among Adults--United States, 2014. *MMWR Morb Mortal Wkly Rep* **2016**, *65*, 137-141, doi:10.15585/mmwr.mm6506a1.
3. Shan, Z.; Ma, H.; Xie, M.; Yan, P.; Guo, Y.; Bao, W.; Rong, Y.; Jackson, C.L.; Hu, F.B.; Liu, L. Sleep duration and risk of type 2 diabetes: a meta-analysis of prospective studies. *Diabetes Care* **2015**, *38*, 529-537, doi:10.2337/dc14-2073.
4. Huang, T.; Redline, S. Cross-sectional and Prospective Associations of Actigraphy-Assessed Sleep Regularity With Metabolic Abnormalities: The Multi-Ethnic Study of Atherosclerosis. *Diabetes Care* **2019**, *42*, 1422-1429, doi:10.2337/dc19-0596.
5. Sanchez-de-la-Torre, M.; Campos-Rodriguez, F.; Barbe, F. Obstructive sleep apnoea and cardiovascular disease. *Lancet Respir Med* **2013**, *1*, 61-72, doi:10.1016/S2213-2600(12)70051-6.
6. Liu, R.Q.; Qian, Z.; Trevathan, E.; Chang, J.J.; Zelicoff, A.; Hao, Y.T.; Lin, S.; Dong, G.H. Poor sleep quality associated with high risk of hypertension and elevated blood pressure in China: results from a large population-based study. *Hypertens Res* **2016**, *39*, 54-59, doi:10.1038/hr.2015.98.
7. Chastin, S.F.; Palarea-Albaladejo, J.; Dontje, M.L.; Skelton, D.A. Combined Effects of Time Spent in Physical Activity, Sedentary Behaviors and Sleep on Obesity and Cardio-Metabolic Health Markers: A Novel Compositional Data Analysis Approach. *PLoS One* **2015**, *10*, e0139984, doi:10.1371/journal.pone.0139984.
8. Alvaro, P.K.; Roberts, R.M.; Harris, J.K. A Systematic Review Assessing Bidirectionality between Sleep Disturbances, Anxiety, and Depression. *Sleep* **2013**, *36*, 1059-1068, doi:10.5665/sleep.2810.
9. Hafner, M.; Stepanek, M.; Taylor, J.; Troxel, W.M.; van Stolk, C. Why Sleep Matters-The Economic Costs of Insufficient Sleep: A Cross-Country Comparative Analysis. *Rand Health Q* **2017**, *6*, 11-11.
10. O'CONNOR, C.; THORNLEY, K.S.; HANLY, P.J. Gender Differences in the Polysomnographic Features of Obstructive Sleep Apnea. *American Journal of Respiratory and Critical Care Medicine* **2000**, *161*, 1465-1472, doi:10.1164/ajrccm.161.5.9904121.
11. Redline, S.; Tishler, P.V.; Hans, M.G.; Tosteson, T.D.; Strohl, K.P.; Spry, K. Racial differences in sleep-disordered breathing in African-Americans and Caucasians. *American Journal of Respiratory and Critical Care Medicine* **1997**, *155*, 186-192, doi:10.1164/ajrccm.155.1.9001310.
12. Kiley, J.P.; Twery, M.J.; Gibbons, G.H. The National Center on Sleep Disorders Research—progress and promise. *Sleep* **2019**, *42*, doi:10.1093/sleep/zsz105.

13. Newell, J.; Mairesse, O.; Verbanck, P.; Neu, D. Is a one-night stay in the lab really enough to conclude? First-night effect and night-to-night variability in polysomnographic recordings among different clinical population samples. *Psychiatry Res* **2012**, *200*, 795-801, doi:10.1016/j.psychres.2012.07.045.
14. Means, M.K.; Edinger, J.D.; Glenn, D.M.; Fins, A.I. Accuracy of sleep perceptions among insomnia sufferers and normal sleepers. *Sleep Med* **2003**, *4*, 285-296, doi:10.1016/s1389-9457(03)00057-1.
15. Fry, J.M.; DiPhillipo, M.A.; Curran, K.; Goldberg, R.; Baran, A.S. Full Polysomnography in the Home. *Sleep* **1998**, *21*, 635-642, doi:10.1093/sleep/21.6.635.
16. Peake, J.M.; Kerr, G.; Sullivan, J.P. A Critical Review of Consumer Wearables, Mobile Applications, and Equipment for Providing Biofeedback, Monitoring Stress, and Sleep in Physically Active Populations. *Frontiers in Physiology* **2018**, *9*, doi:10.3389/fphys.2018.00743.
17. Bianchi, M.T. Sleep devices: wearables and nearables, informational and interventional, consumer and clinical. *Metabolism* **2018**, *84*, 99-108, doi:10.1016/j.metabol.2017.10.008.
18. Rosvold, H.E.; Mirsky, A.F.; Sarason, I.; Bransome Jr, E.D.; Beck, L.H. A continuous performance test of brain damage. *Journal of consulting psychology* **1956**, *20*, 343.
19. de Zambotti, M.; Claudatos, S.; Inkelis, S.; Colrain, I.M.; Baker, F.C. Evaluation of a consumer fitness-tracking device to assess sleep in adults. *Chronobiology International* **2015**, *32*, 1024-1028, doi:10.3109/07420528.2015.1054395.
20. de Zambotti, M.; Rosas, L.; Colrain, I.M.; Baker, F.C. The Sleep of the Ring: Comparison of the ŌURA Sleep Tracker Against Polysomnography. *Behavioral Sleep Medicine* **2019**, *17*, 124-136, doi:10.1080/15402002.2017.1300587.
21. 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* **2015**, *38*, 1323-1330, doi:10.5665/sleep.4918.
22. Montgomery-Downs, H.E.; Insana, S.P.; Bond, J.A. Movement toward a novel activity monitoring device. *Sleep and Breathing* **2012**, *16*, 913-917, doi:10.1007/s11325-011-0585-y.
23. Liu, W.; Ploderer, B.; Hoang, T. In Bed with Technology: Challenges and Opportunities for Sleep Tracking. In Proceedings of Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, Parkville, VIC, Australia; pp. 142-151.
24. Piwek, L.; Ellis, D.A.; Andrews, S.; Joinson, A. The Rise of Consumer Health Wearables: Promises and Barriers. *PLOS Medicine* **2016**, *13*, e1001953, doi:10.1371/journal.pmed.1001953.
25. Murakami, H.; Kawakami, R.; Nakae, S.; Nakata, Y.; Ishikawa-Takata, K.; Tanaka, S.; Miyachi, M. Accuracy of Wearable Devices for Estimating Total Energy Expenditure: Comparison With Metabolic Chamber and Doubly Labeled Water Method. *JAMA Internal Medicine* **2016**, *176*, 702-703, doi:10.1001/jamainternmed.2016.0152.
26. Xie, J.; Wen, D.; Liang, L.; Jia, Y.; Gao, L.; Lei, J. Evaluating the Validity of Current Mainstream Wearable Devices in Fitness Tracking Under Various Physical Activities: Comparative Study. *JMIR Mhealth Uhealth* **2018**, *6*, e94, doi:10.2196/mhealth.9754.
27. Lee, J.-M.; Byun, W.; Keill, A.; Dinkel, D.; Seo, Y. Comparison of Wearable Trackers' Ability to Estimate Sleep. *International Journal of Environmental Research and Public Health* **2018**, *15*, 1265.
28. de Zambotti, M.; Goldstone, A.; Claudatos, S.; Colrain, I.M.; Baker, F.C. A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiology International* **2018**, *35*, 465-476, doi:10.1080/07420528.2017.1413578.

29. Buysse, D.J.; Hall, M.L.; Strollo, P.J.; Kamarck, T.W.; Owens, J.; Lee, L.; Reis, S.E.; Matthews, K.A. Relationships between the Pittsburgh Sleep Quality Index (PSQI), Epworth Sleepiness Scale (ESS), and clinical/polysomnographic measures in a community sample. *J Clin Sleep Med* **2008**, *4*, 563-571.

30. Kane, M.J.; Conway, A.R.A.; Miura, T.K.; Colflesh, G.J.H. Working memory, attention control, and the n-back task: A question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **2007**, *33*, 615-622, doi:10.1037/0278-7393.33.3.615.

31. Killgore, W.D.S. Effects of sleep deprivation on cognition. In *Progress in Brain Research*, Kerkhof, G.A., Dongen, H.P.A.v., Eds. Elsevier: 2010; Vol. 185, pp. 105-129.

32. Perez, M.V.; Mahaffey, K.W.; Hedlin, H.; Rumsfeld, J.S.; Garcia, A.; Ferris, T.; Balasubramanian, V.; Russo, A.M.; Rajmane, A.; Cheung, L. Large-scale assessment of a smartwatch to identify atrial fibrillation. *New England Journal of Medicine* **2019**, *381*, 1909-1917.

33. Miotto, R.; Danieletto, M.; Scelza, J.R.; Kidd, B.A.; Dudley, J.T. Reflecting health: smart mirrors for personalized medicine. *NPJ digital medicine* **2018**, *1*, 1-7.