Improved Sleep Detection Through the Fusion of Phone Agent and Wearable Data Streams

Gonzalo J. Martinez[1*], Stephen M. Mattingly[1*], Jessica Young[1], Louis Faust[1],
Anind K. Dey[2], Andrew T. Campbell[3], Munmun De Choudhury[4], Shayan Mirjafari[3],
Subigya K. Nepal[3], Pablo Robles-Granda[1], Koustuv Saha[4], Aaron D. Striegel[1]
[1]University of Notre Dame, [2]Dartmouth College, [3]University of Washington
[1]{gmarti11, smattin1, jyoung22, lfaust, problesg, striegel}@nd.edu, [2]anind@uw.edu,
[3]{shayan, sknepal, campbell}@cs.dartmouth.edu, [4]{munmund, koustuv.saha}@gatech.edu,

***Abstract*—Commercial grade activity trackers and phone agents are increasingly being deployed as sensors for sleep in large scale, longitudinal designs. In general, wearables detect sleep through diminished movement and decreased heart rate (HR), while phone agents look for lack of user input, or lack of movement, sound or light. However, recent literature suggests that commercial-grade wearables and phone apps vary greatly in the accuracy of sleep predictions, which along the constant innovation in each new version and proprietary algorithms, make it difficult to evaluate their efficacy for scientific study, especially outside of the laboratory. In a longitudinal study, we find that wearables cannot detect when a person is laying still but using their phones, a common behavior, overestimating sleep when compared to self-reports. Therefore, we propose that fusing these sensors allows for more accurate sleep detection by capitalizing on the benefits of both streams: combining the movement detection of wearables with the technology usage detected by cell phones. We determine that fusing phone activity to wearables can generate better models of self-reported sleep than either stream alone, testing these models in two separate datasets.**

***Index Terms*—wearables; Phone; Sensor fusion; sleep**

I. Introduction

Commercial grade activity trackers as well as phone agents have seen an uptick in deployment as sensors for sleep in large scale, longitudinal designs [1]–[3]. These complement previous large scale and epidemiological studies that find relationships between self-report and health e.g., [4]. However, recent literature suggests that commercial grade wearables and phone apps vary greatly in accuracy of sleep measurements, and validation is made difficult as wearables and apps are frequently updated and modified. However, wearables usually work by intergrating a photoplethysmogram (PPG) sensor used to detect HR and an accelerometer used to detect the lack of movement or patterns of movement associated with sleep. Phone apps, on the other hand, can make use of the sound sensor, light sensor, accelerometer, as well as phone usage to detect sleep [5]. Given how these two sensors measure two different aspects, one capturing a signal straight from the body while the other captures interaction with the environment, we set out to determine if fusing phone activity to wearables can generate better models of sleep and self report than either stream alone. These technologies can complement each other through combined assessment of the body (wearable) and conscious use of a phone while stationary (phone agent).

*Authors contributed equally to this work

A. Wearable Agreement with Gold Standard

While many commercial devices have been studied, the relationship between the gold standard of sleep, Polysomnography (PSG), and all other measures (actigraphy, wearables, phone agent, sleep dairy, self-report) varies, with strong agreements in lab and significantly weaker relationships *in situ*. For sleep measurements using phone usage, actigraphy, and wearables which rely on patterns of movement to determine sleep, variability logically can stem from either lack of motion while awake (e.g., watching a movie, using a phone in bed), or movement not associated with a person's sleep (e.g., pet, bed partner) [6]. These real-world situations may help explain why sleep trackers may do fairly well in the laboratory (e.g., within 9 minutes for actigraphy and PSG) [7] and fairly poorly outside of the laboratory if independent validation of devices is even available [8].

Further, researchers have seen differences in accuracy of sleep metrics comparing wearable technology to PSG, depending on study population (healthy, or with insomnia [9], major depressive disorder [10], age of population [11], [12], etc.). In general, actigraphy and wearables tend to accurately reflect PSG in laboratory studies of healthy populations. For instance, [9] used the Fitbit Flex and PSG to see how accurately the wearable tracked the sleep of healthy sleepers and insomniacs. The wearable device showed strong agreement with PSG. [9] reported the Fitbit Flex in "normal mode"overestimated Total Sleep Time (TST) relative to PSG by 6.5 mins for healthy sleepers, and was off by 32.9 minutes for insomniacs. [7] used 100 participants from the general population, aged 18-75 from one town, and found that a research grade actigraph underestimated sleep by 8.3 minutes relative to PSG when worn on the wrist. As another example, [10] also used the Fitbit Flex and a research grade actigraphy watch for people with Major Depressive Disorder, and found that compared to PSG the Fitbit Flex overestimated TST by an average of 46 minutes. In "sensitive mode", a mode of operation that shows more details about sleep than the "normal mode", the Fitbit Flex underestimated TST by an average of 86.3 minutes. Similar to the wearable in normal mode, the research grade

actigraph overestimated sleep by 40.6 minutes. In addition to variability around sleep accuracy, wearables also may suffer from significant missing data (e.g., [13].

B. Mobile App and Sleep Detection

Smartphones have many datastreams available and as such have been used to detect sleep [5], [14]. As one example of phone usage and self-reported sleep, [15] collected a sleep diary from 27 participants. After collecting training data for 3 weeks, their app was able to get within 50 minutes of sleep diaries. For objective measures, [16] used 400 participants to build and test models using cell phone usage compared to a commercial wearable and obtained 89% agreement with the commercial wearable. Comparing mobile measurement to PSG, [17] was able to get 89% agreement between the phone sleep measuring app and PSG from 20 participants. In [16] sleep duration statistics were not reported per-se and the work noted that performance is likely to be worse in detecting sleep in insomniacs.

C. Self-report and Objective Sleep

In a large study, [18] used over 2,000 participants with a mean age of 67 years and a standard deviation of 10 years, and showed that participants overestimated sleep duration after a PSG recorded night of sleep by a statistically significant 16 minutes. When asked about their habitual sleep time relative to the night of PSG, they overestimated by 59 minutes. These specific differences varied slightly by age, race, obesity status, education, and sex. For instance, men overestimated sleep by 24 minutes while women overestimated sleep by 10 minutes. Even in clinical patients, the agreement between self-reported sleep duration and PSG assessed sleep is fairly good, with a difference of less than 10 minutes on average across 101 participants with various sleep disorders, which was not statistically significantly different relative to PSG [19]. However, not all studies show as strong agreement as above. For instance, [11] used self-report and actigraphy and suggested that self reports and actigraphy correlate, but did not agree in a pool of 225 adolescent participants. Specifically, self-reports overestimate actigraphy by an average of 95.5 minutes during the week and 156.7 minutes during the weekend. In addition, [7] reported that while actigraphy underestimated sleep by 8.3 minutes when worn on the wrist, self-reported measures underestimated by 39.8 minutes.

D. Summary

In summary, significant variability exists between self-reported measures of sleep duration and objectively assessed measures of sleep duration, though subjective measures still remain relevant for health. For a discussion about the utility and drawbacks of self-reported sleep as an epidemiological tool, see [20], [21]. Within objectively measured sleep duration, there is additional variability between smartphone, wearable, actigraphy, and PSG assessed sleep. While performance between various mobile apps and wearables varies depending on device and assessment method, most papers reviewed here show less than an hour of discrepancies across all measurement modalities. In this paper, we address the discrepancies between self-reported sleep as assessed by mobile app and self-reported sleep as assessed by wearable by fusing wearable and mobile app data together. By combining the measures, we hope to get a better measure than when using either device alone.

II. Methods

Data was collected from the Tesserae study [1], and included wearable data, phone usage data, and daily survey responses. Study details are fully described in [1].The protocol was fully approved by the Institutional Review Board. For this paper, we reiterate relevant data characteristics and streams.

A. Participants

The Tesserae Study [1] recruited 757 participants from cognitively demanding professions (e.g., information workers) for a year-long study. Participants were recruited throughout the United States, concentrated around four major organizations. Participants provided data from commercially available sensors (e.g., wearable, smartphone tracker - see below), and this data was used to predict several behavioral and psychological constructs, including daily sleep duration. We consider data from 575 as due to project funding requirements, 151 were blinded (data was withheld from researchers to validate researcher work). Furthermore, 19 participants did not provide demographic information, 11 participants dropped before daily survey completion, and 1 participant did not answer any daily surveys. Table I provides demographic information for the study participants.

B. Wearable

Participants received a Garmin Vivosmart 3 with the request to wear the device at all times excluding showering and charging. The wearable was chosen, in part, for an approximate 5 day battery life, rapid charging capability, and HR streaming capabilities.

C. Phone Agent

Using an app based on StudentLife [5], we collected usage data such as screen locks and unlocks from both iOS and Android models. This phone agent, (PA), worked differently for iOS and Android; iOS models sampled screen state every 10 minutes to conserve battery while Android provided time stamped screen on and off times, which may result in slight differences in values between platform.

D. Ground Truth Survey

For the first 56 days, participants received a daily Qualtrics survey via text message at 8am, noon, or 4pm, and had 4 hours to complete. Several versions of the daily text message asked about a variety of constructs; here, we consider the question "How many hours of actual sleep did you get LAST NIGHT?". This question was asked 3-4 times per week and at either 8am or noon relative to the timezone of the participant, having a maximum of 28 answers per participant.

E. Calculations for Sensor Fusion

In the case of the wearable, sleep duration was calculated from the bed and wake times provided from the Garmin Health API. From the phone agent, bedtime was defined as the last detected conscious action of the evening and the first conscious detected action of the morning (e.g., screen unlocking). We formed a combined measure of sleep duration by starting with the wearable bed and wake times and by examining phone agent activity up to 90 minutes after wearable detected bed time, and 90 minutes prior to wearable detected wake time. These windows were chosen based on our literature review which suggested the largest reported error of a wearable and other methods of sleep detection was $\sim$86 minutes [10] , and it was not clear whether this error occurred as a result of poor detection of bed time, wake time, a combination of both, or detection of significant periods of wake time during sleep across the whole night. If conscious phone use (defined as a screen unlock) was detected within 90 minutes of bed and/or wake, the sleep duration was adjusted accordingly to the last detected unlock after bed time and the first detected unlock before wake time. For instance, a wearable detected bed time of 9:30pm and a wake time of 6:30am. Phone activity was detected at 9:45pm, 10:00pm, 2:00am, and 6:15am. After applying our fusion method, we would have a combined bed time of 10:00pm and a wake time of 6:15am, which would result in an adjusted sleep duration of 8.25 hours from a wearable that detected 9 hours.

F. Calculation for Random Baseline Condition

In order to ensure that the combined measure is better than randomly reducing sleep duration, we generated a random baseline calculation. In total, ten measures were created by taking the sleep value reported by the wearable and adjusting it by a random amount between 1 and 180 minutes, the minimum and maximum adjustments our combined measure could receive. For the results, we only report the best scoring random model, rather than all 10 iterations.

G. Calculation for Combined and Single Sensor Condition

We anticipate that requiring all sensor streams, demographics, and daily responses will result in a significant reduction in usable data. In an effort to increase coverage of the available ground truth data and take advantage of multiple sensor streams, we generated a sleep duration combined measure which uses combined sensors where available and single streams when combined streams are not available. We applied the following logic: If wearable and phone agent are available, combine as described above using fusion method. If only a single stream was available, use that (either wearable only or phone agent only), or use the mean of the measured data up to that point.

III. Results

From a theoretical maximum response rate of 16,100 (575 participants x 28 sleep surveys). 15,582 days had survey responses (96.7%). 12,789 days had data for at least one of the two sensors. From this, we have 9,127 datapoints from 525 participants that have a self report, demographics, a wearable sleep duration, and phone agent data with sufficient usage to generate a sleep duration (e.g., a last daily unlock and first morning unlock) that we use for comparison between the sensors by themselves vs our sensor fusion method. However, as mentioned in the previous section, we also followed a combined and single sensor approach which will use the entirety of 12,789 days that have at least one sensor available.

Demographics

Construct		Count		Count
Sex	Male	301	Female	224
Platform	iOS	311	Android	214
Role	Supervisor	239	Non-supervisor	286
Age	Mean	34.7	Standard deviation	9.6

Income	**Count**
>$25,000	3
25,000−49,999	39
50,000−74,999	112
75,000−99,999	113
100,000−124,999	93
125,000−150,000	52
<$150,000	113

TABLE I
DEMOGRAPHIC INFORMATION

A. Descriptives

From the survey, we found a self-reported average sleep duration of 7.03± .01 hours. Wearable's average bed time was 11:07pm ± <1 minute, wake time averaged 6:55am ± <1 minute with an average duration of 7.8 hours± .02 hours. From the phone agent, we observed an average bed time of 11:22pm ± <1 minute , an average wake time of 6:52am ± <1 minute , and an average duration of 7.48 hours of sleep,± .02 hours.

We made 1585 adjustments to bed time, average adjustment of 44.3 minutes, ± 1 minute, 2622 adjustments to wake time, with an average adjustment of 30.7 minutes to wake, ± 1 minute, 2618 adjustments to both; with an average adjustment of 77.19 minutes and± 1 minute, The combined measure had an average bed time of 11:27pm ± 1 minute, an average wake time of 6:37am± 1 minute , and an average duration of 7.16 ± .018 hours of sleep. for a breakdown of sleep adjustments, see figure 1.

B. Model Comparisons

We tested our adjusted measure by comparing each measure, wearable sleep duration, phone agent sleep duration, randomly adjusted sleep duration, and our combined sleep duration measure, with self-reported sleep to obtain MAE and RMSE. Additionally, given the repeated measures design, we ran linear mixed effects models with subject as a random effect. We included demographic information including sex, platform, age, supervisor status, and income as fixed factor predictors as well as a sleep duration measure as predictor; either wearable, phone agent, randomly adjusted sleep duration, or our

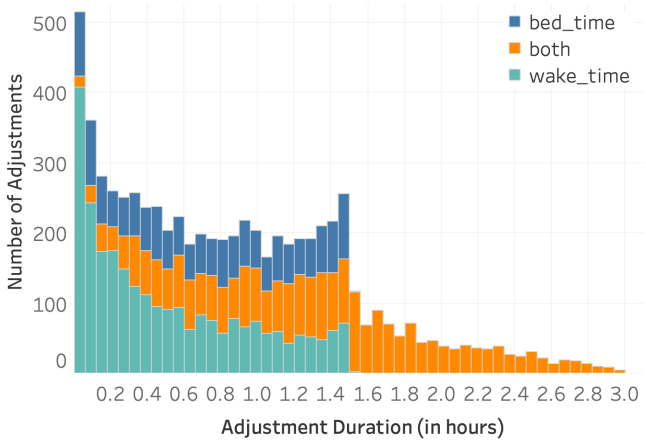

Fig. 1. Histogram of sleep adjustment duration in hours. Light blue is wake adjustment only, Dark blue is bedtime adjustment only, and orange is both bed and wake time adjustments

combined sleep duration measure. As expected, the combined feature performed best as noted by the lowest AIC, BIC, MAE, and RMSE (see table II). When looking at average sleep duration, (see Figure 2), the combined measure is only 8 minutes off of self-reported sleep, compared to 27 minutes for phone agent only and 46 minutes for wearable only.

	Meas. + Dem.		Measure	
Predictor	*AIC*	*BIC*	*MAE*	*RMSE*
Best Random	29387	29486	1.59	2.17
Wearable	29169	29268	1.38	2.00
Phone Agent	29047	29147	1.32	1.87
Combined	29010	29109	1.15	1.76

TABLE II
TABLE 2. A SUMMARY OF MODEL AND MEASURE PERFORMANCE FOR PREDICTING SELF-REPORTED SLEEP

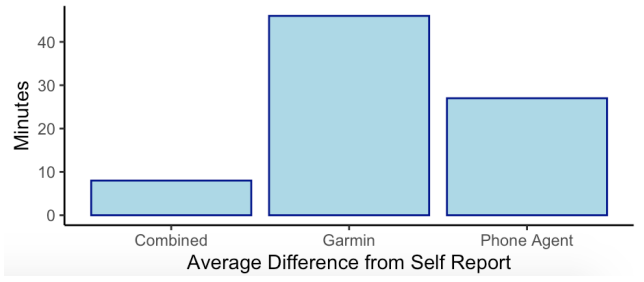

Fig. 2. Comparisons of sleep duration.

C. Combined and Single Sensor

We examined the efficacy of our dual stream approach. In our dataset we have 12,789 points where the wearable data was available, 10,782 where phone agent was available, 9,253 data points with both sensors available, and 1,264 points where none were available. From this, we constructed an extended dataset with 12,789 data points with at least one data stream. This amounts to 3,662 more data points available (approximately 20%) relative to our complete data set (no missing data, no imputation).

When considering the extended dataset, we find an RMSE of 1.87 for the adjusted measure that also uses a single stream when both are not available. This is better than a baseline mean imputation for the wearable, RMSE=2.05, or using phone agent and the wearable (without combining them when both available) RMSE=1.98. When comparing the measures for the complete dataset and the extended dataset, we find an an RMSE of 1.76 for the complete dataset and an RMSE of 1.87 for the extended dataset. This difference of RMSE of .11 translates into about 6.6 minutes of additional error.

D. Exploratory: Who is being adjusted

After establishing that the adjusted sleep value is most effective for predicting self-reported sleep duration, we explored how this adjustment was generated. Specifically, we sought to determine if many people had small adjustments, or a few people had larger adjustments. We generated a histogram from the percent of days adjusted out of total days collected for each person (Figure 3), and also generated a histogram from the average duration of correction for each person (Figure 4). Figure 3 is right skewed, suggesting that the majority of participants received many adjustments, and Figure 4 shows that the average adjustment per participant is normally distributed. From these plots we conclude that adjustments are distributed across many individuals for a modest amount, rather than concentrated with large adjustments for a few participants. In addition, we examined what demographic features were associated with the adjustment duration and adjustment frequency.

We ran step-wise regression models including sex, platform, age, supervisor status, and income as predictors. For duration, a significant model was found, $F(1, 519) = 55.087$, $p = < .001$, with an adjusted $R^2 = .096$. In the final model, participant's duration of adjustment in hours = .995 -.225 platform, with 0 coded as Android and 1 coded as iOS. Thus, the average adjustment for iOS duration was 0.77$\pm$ .02 hours, while for Android the average adjustment was .99$\pm$ .03 hours. We repeated the step-wise regression for frequency of adjustments as the dependent variable, with a significant model found $F(3, 517) = 16.179$, $p < .001$, with this adjusted $R^2 = .081$. The model predicted number of adjustments out of total days collected as .948% -.099 for platform (0 = Android, 1 is iOS) -.004(age, in years), + .058 supervisor status (0 = non-supervisor, supervisor = 1). Thus, Android users, younger participants, and supervisors had more phone agent adjustments on average.

E. Exploratory: Adjustments on a different dataset

Now that we have demonstrated that adjusting wearable bed and wake times by using a phone agent is a viable approach, we explored phone usage and wearable bed and wake times in a different population (students), and with a different device (Fitbit Charge HR 2), using data from the Net-Health project

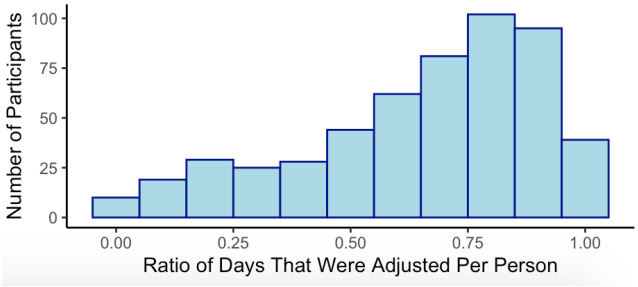

Fig. 3. Histogram of ratio of data adjusted within person.

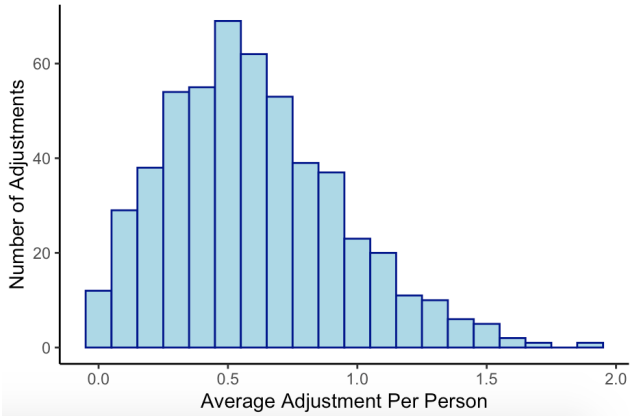

Fig. 4. Histogram of average sleep adjustment duration per person.

[2]. This dataset contains data from 486 students for four years, for a total of 129,858 days of wearable and phone usage data. This dataset does not contain self-reports of sleep. However, we applied our approach to corroborate whether the adjustments would happen in a different population, that of college of students, to better understand the generalizability of our method. We adjusted 19,929 bed times with an average adjustment of 42 $\pm$ .1 minutes, 28,546 adjustments to wake up time with an average of 31 minutes $\pm$ .1 minutes, and 7,565 adjustments to both bed and wake time, for an average of 78 $\pm$ .5 minutes. After adjustments, we found an average sleep duration of 7.35 hours $\pm$ 1 minute, with an average bed time of 1:27am $\pm$ 2 minutes, and an average wake time of 8:49am $\pm$ 2 minutes.

IV. Discussion

As sensors such as wearables and smartphones become more pervasive, the utility of features derived from multiple streams will increase. Specifically, combined sensors allow us to improve sleep estimates from the different behaviors that impact sleep measurement (actigraphy, phone usage), and second, to generate data from an alternative stream which would be missing in the case of a single stream.

Here, we demonstrated improved agreement between subjective sleep experience by combining the actigraphy/PPG based wearable sleep detection with the behavioral sensing of phone usage after bed time and before wake time. While these sensors do not cover all cases in which wearable sleep sensing and subjective sleep experiences may diverge (e.g., reading a book in bed), we did adjust $\sim$75% of all wearable measurements with phone usage data. after adjustment, we were within 8 minutes of self-reported sleep.

One downside to a multi-stream approach is compliance/maintenance. When requiring both sensor streams instead of a single sensor stream, we reduced the number of samples of our dataset by over 40%. This also makes data collection tricky, as different streams have different requirements and different life cycles, especially in longitudinal studies. For instance, a change in iOS or Android may reduce the phone agent's utility, while at a different time an update from a wearable manufacturer may change how data is collected, or a participant's phone may break while the wearable continues to function, etc. However, we proposed and demonstrated a method in which dual streams and single streams, along with the union for imputed values, can recover a significant portion of missing data. With this method, accuracy was only slightly eroded (within $\sim$ 14 minutes of self-reported sleep) when the sensor fusion also allowed single stream and imputed values, and this allowed for recovery of 20% of data. Thus, our combined sleep measure is accurate in both complete datasets and in situations with missing data from single or both streams, and using multiple datastreams can offset missing data in a single stream reasonably well.

We also demonstrated that applying our method to a different dataset, with different wearables, apps for assessing phone usage, and a different population yields nearly identical adjustment amounts; wakeup time within 1 minute, bed time within 2 minutes, and both wakeup and bed time within 1 minute. However, the percentages of adjustments were different, with Tesserae participants having $\sim$75% of values adjusted, while the Net-Health dataset was adjusted only $\sim$43% of the time. These two datasets demonstrate the robustness of our approach as far as total duration in relation to wearable, phone usage sensor, and population, while the differences in adjustment percentage suggest that these populations may have different phone usage and sleep behaviors.

Finally, we examined what demographic features affected the adjustment and how the adjustments were distributed. We report models for duration and frequency with R^2s less than .1. within these models, for the amount of adjustment differs only by platform, with differences that translate to less than 14 minutes. These differences may be explained by the differences in detecting screen usage in iOS, which samples every 10 minutes, and Android, which timestamps each lock and unlock of the PA. For frequency, again we see modest differences between iOS and Android, and additional effects of age and supervisor status. When looking at the distribution of frequency and adjustment amounts, we found that most participants received adjustments for a modest amount, rather than concentrated in a few users who use phones for long durations. Taken all together, this suggests that our combined feature may be reasonably applied to and useful for broad populations.

V. Conclusion

Combining meaningful sensors streams can greatly improve the accuracy of wearable measurements relative to self report. This level of accuracy might allow future researchers to replace self-reported EMAs in future studies. In addition, given that many wearables already pair with phones and/or utilize apps (e.g., Garmin Connect), wearable manufacturers could improve device accuracy by including phone usage in sleep detection algorithms.

VI. Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2017- 17042800007. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] S. M. Mattingly and e. al., "The Tesserae Project: Large-Scale, Longitudinal, *In Situ,* Multimodal Sensing of Information Workers," in *Extended Abstracts of CHI '19*. Glasgow, Scotland Uk: ACM Press, 2019, pp. 1–8.

[2] R. Purta, S. Mattingly, L. Song, O. Lizardo, D. Hachen, C. Poellabauer, and A. Striegel, "Experiences measuring sleep and physical activity patterns across a large college cohort with fitbits," in *Proc. of ISWC'16*, 2016, pp. 28–35.

[3] K. G. Baron, J. Duffecy, M. A. Berendsen, I. Cheung Mason, E. G. Lattie, and N. C. Manalo, "Feeling validated yet? A scoping review of the use of consumer-targeted wearable and mobile technology to measure and improve sleep," *Sleep Medicine Reviews*, vol. 40, pp. 151–159, Aug. 2018.

[4] E. Kronholm, T. Laatikainen, M. Peltonen, R. Sippola, and T. Partonen, "Self-reported sleep duration, all-cause mortality, cardiovascular mortality and morbidity in Finland," *Sleep Medicine*, vol. 12, no. 3, pp. 215–221, Mar. 2011.

[5] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell, "StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones," ser. UbiComp '14. New York, NY, USA: ACM, 2014, pp. 3–14.

[6] D. Aliakseyeu, J. Du, E. Zwartkruis-Pelgrim, and S. Subramanian, "Exploring Interaction Strategies in the Context of Sleep," in *Human-Computer Interaction – INTERACT 2011*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, vol. 6948, pp. 19–36.

[7] M. Zinkhan, K. Berger, S. Hense, M. Nagel, A. Obst, B. Koch, T. Penzel, I. Fietze, W. Ahrens, P. Young, S. Happe, J. W. Kantelhardt, A. Kluttig, A. Schmidt-Pokrzywniak, F. Pillmann, and A. Stang, "Agreement of different methods for assessing sleep characteristics: a comparison of two actigraphs, wrist and hip placement, and self-report with polysomnography," *Sleep Medicine*, vol. 15, no. 9, pp. 1107–1114, Sep. 2014.

[8] J. M. Peake, G. Kerr, and J. P. Sullivan, "A Critical Review of Consumer Wearables, Mobile Applications, and Equipment for Providing Biofeedback, Monitoring Stress, and Sleep in Physically Active Populations," *Frontiers in Physiology*, vol. 9, Jun. 2018.

[9] S.-G. Kang, J. M. Kang, K.-P. Ko, S.-C. Park, S. Mariani, and J. Weng, "Validity of a commercial wearable sleep tracker in adult insomnia disorder patients and good sleepers," *Journal of Psychosomatic Research*, vol. 97, pp. 38–44, Jun. 2017.

[10] J. D. Cook, M. L. Prairie, and D. T. Plante, "Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: A comparison against polysomnography and wrist-worn actigraphy," *Journal of Affective Disorders*, vol. 217, pp. 299–305, Aug. 2017.

[11] T. Arora, E. Broglia, D. Pushpakumar, T. Lodhi, and S. Taheri, "An Investigation into the Strength of the Association and Agreement Levels between Subjective and Objective Sleep Duration in Adolescents," *PLoS ONE*, vol. 8, no. 8, p. e72406, Aug. 2013.

[12] M. de Zambotti, F. C. Baker, and I. M. Colrain, "Validation of Sleep-Tracking Technology Compared with Polysomnography in Adolescents," *Sleep*, vol. 38, no. 9, pp. 1461–1468, Sep. 2015.

[13] A. Baroni, J.-M. Bruzzese, C. A. Di Bartolo, and J. P. Shatkin, "Fitbit Flex: an unreliable device for longitudinal sleep measures in a non-clinical population," *Sleep and Breathing*, vol. 20, no. 2, pp. 853–854, May 2016.

[14] Z. Chen, M. Lin, F. Chen, N. D. Lane, G. Cardone, R. Wang, T. Li, Y. Chen, T. Choudhury, and A. T. Campbell, "Unobtrusive Sleep Monitoring Using Smartphones," ser. PervasiveHealth '13. ICST, Brussels, Belgium, Belgium: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2013, pp. 145–152.

[15] J.-K. Min, A. Doryab, J. Wiese, S. Amini, J. Zimmerman, and J. I. Hong, "Toss 'n' turn: smartphone as sleep and sleep quality detector," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. Toronto, Ontario, Canada: ACM Press, 2014, pp. 477–486.

[16] A. Cuttone, P. Bækgaard, V. Sekara, H. Jonsson, J. E. Larsen, and S. Lehmann, "SensibleSleep: A Bayesian Model for Learning Sleep Patterns from Smartphone Events," *PLOS ONE*, vol. 12, no. 1, p. e0169901, Jan. 2017.

[17] S. Bhat, A. Ferraris, D. Gupta, M. Mozafarian, V. A. DeBari, N. Gushway-Henry, S. P. Gowda, P. G. Polos, M. Rubinstein, H. Seidu, and S. Chokroverty, "Is There a Clinical Role For Smartphone Sleep Apps? Comparison of Sleep Cycle Detection by a Smartphone Application to Polysomnography," *Journal of Clinical Sleep Medicine*, Jul. 2015.

[18] G. E. Silva and J. A. Walsleben, "Relationship Between Reported and Measured Sleep Times: The Sleep Heart Health Study (SHHS)," vol. 3, no. 6, p. 9, 2007.

[19] P. A. Vanable, J. E. Aikens, L. Tadimeti, B. Caruana-Montaldo, and W. B. Mendelson, "Sleep Latency and Duration Estimates Among Sleep Disorder Patients: Variability as a Function of Sleep Disorder Diagnosis, Sleep History, and Psychological Characteristics," *Sleep*, vol. 23, no. 1, pp. 1–9, Jan. 2000.

[20] M. T. Bianchi, R. J. Thomas, and M. B. Westover, "An open request to epidemiologists: please stop querying self-reported sleep duration," *Sleep Medicine*, vol. 35, pp. 92–93, Jul. 2017.

[21] B. Kowall and A. Stang, "Measurement is always better than self-report: is it that easy?" *Sleep Medicine*, vol. 38, p. 157, Oct. 2017.