On the Quality of Real-world Wearable Data in a Longitudinal Study of Information Workers

Gonzalo J. Martinez[1], Stephen M. Mattingly[1], Shayan Mirjafari[2], Subigya K. Nepal[2],
Andrew T. Campbell[2], Anind K. Dey[3], Aaron D. Striegel[1]
[1]University of Notre Dame, [2]Dartmouth College, [3]University of Washington
[1]{gmarti11,smattin1,striegel}@nd.edu, [2]{shayan, sknepal,campbell}@cs.dartmouth.edu, [3]anind@uw.edu

***Abstract*—Over the past few years, an incredible diversity of consumer-grade wearables has emerged with tremendous breadth in capabilities, form factor, and cost. These wearable devices show significant promise for researchers to conduct expansive research studies in terms of scale, scope, and duration. Unfortunately, there is limited public data shared with respect to the data quality, device longevity, and scaling issues that emerge when trying to execute such studies. To that end, we share real-world results with respect to data quality, participant compliance, and device efficacy on a large scale, longitudinal study involving over seven hundred and fifty working professionals over the period of an entire year. In this paper, we present analyses with respect to the different types of data being collected including sleep, heart rate, physical activity, and stress. Furthermore, we explore participants behavior regarding charging frequency, and device robustness to further aid researchers considering large scale wearable studies.**

***Index Terms*—Wearables; User Studies; Sleep; Heart Rate, Data Quality**

I. Introduction

Over the past few years, the quality and sophistication of wearable devices have improved dramatically. Wearables have evolved into complex devices that can connect to smartphones and provide notifications, alerts, physical activity, sleep recognition, real time heart rate monitoring, and more. Accompanied by a steady increase in the adoption of these devices [1], the possibilities for their use seem endless, making wearables a veritable treasure trove of opportunities for researchers.

The widespread adoption of these sensor rich consumer grade (oriented) devices means that one can now conduct research at a larger scale in both the size and duration of studies to explore numerous aspects of the human condition. Unfortunately, despite the fact that wearables have become more affordable, it is still expensive to conduct large studies. Furthermore, trying to conduct large studies becomes harder when going beyond the college campus. Various studies [2]–[5] have tackled individual study deficiencies with respect to time, size, and representative populations but, to the best of our knowledge, there are limited studies that have tried to look more broadly with a sizable population of real world professionals over an extended period of time.

As the community tries to embark on larger studies, how much data can we collect and what is the quality and consistency of that data? To that end, we share in this paper our experiences and analysis on data quality with using wearables in conducting such large study. The contributions of this paper are as follows:

- *Analysis of large-scale data consistency:* We examine the wearable data as recorded from 757 working professionals who wore a study-provided Garmin vivoSmart 3 ranging from nine months to a year. We were able to gather meaningful data roughly 73.5% of the time inclusive of any wearable sub-streams with data either largely being recorded in its entirety (heart rate, steps, stress, etc.) or not at all. Compliance peaked at 90% data peaked before 30 days post-enrollment before stabilizing closer to 70% after 190 days in the study.
- *Analysis of data gaps, replacements, and wearable energy levels:* We examine data gaps and find that 77.2% of missing data happen in gaps of greater than 24 hours with week-long gaps accounting for 70.8% of all missing data for individuals who did not drop out of the study.
- *Analysis of wearable accuracy*: we compare wearable heart rate data for 31 participants with a Zephyr Bioharness in different time windows. We find a fair ICC (.44), MAE of at least 8.4 bpm and MAPE of at least 10.4% between the two measurements.

II. Related Work

In the ubiquitous computing community, there has been considerable interest in monitoring nearly all aspects of human behavior through the use of smartphones [6], wearables [2]–[5], social media [7], and numerous other sensing modalities [8], [9]. Indeed, each year seems to bring new advances and opportunities for sensing offering researchers a myriad of options when designing a study. However, the focus of this paper is strictly on the issue of wearables and in particular, the underlying data consistency, the overarching user wearable compliance, and various behavioral patterns of participants during longitudinal studies involving wearables for which there is a relative dearth of information available as studies focus more often on the end results rather than on the underlying data quality.

The most common type of communication for sharing experiences tend to be 'lessons learned' papers which share hard-earned experiences dealing with the quirks of devices, study management, and scale in terms of study size and duration [10], [11]. The diversity in terms of such studies is often quite significant ranging from the order of less than a

hundred participants over a semester such as with StudentLife [4], [6] to studies approaching hundreds [12] if not nearly one thousand [3] participants. Notably, studies such as [12] and this paper stand out by focusing on working professionals whereas the vast majority of work tends to utilize students and nearby family populations for convenience.

From a data quality perspective, the work by Jeong et. al [2] stands out as the work exclusively focused on the data quality from a population of 50 student participants using Apple Watches over 200 days. In contrast to works that have tried to explore issues associated with longitudinal studies [4], to predict compliance [11], or to increase study compliance [10], the work in [2] focused on the question we are most interested in, namely how much data will one be able to collect, how accurate is the wearable, how does data collection vary across the day, and how do aspects such as wearable energy levels interplay to impact data collection. It is this question of the 'goodness' of practical wearable data that leads us to dive into the wearable data performance that we observed in our large scale longitudinal study of working professionals and to share our observations with the research community.

III. Study Overview

The focus of our overarching study dubbed Tesserae [13] was to explore the extent to which various widely available sensing streams could be used to better characterize various aspects of daily life with a particular focus on individuals in knowledge-based professions. Our study was constructed to leverage a variety of sensing streams including a wearable as in [3], a smartphone sensing agent inspired by [6], Bluetooth LE beacons inspired by [9], and social media analyses [7]. At the close of enrollment which concluded during the Summer of 2018, a total of 757 participants were enrolled in the study,590 of them coming from four major organizations and the rest coming from nearby organizations around the communities of the major organizations. Participants were provided with a Garmin vivoSmart 3 wearable, a phone agent, beacons, and were asked to complete daily surveys. Participants were also asked to provide read-only access to social media, although this last step was not required for participation. In turn, participants were compensated through stipends and / or lottery draws subject to the preference of the employer. Participants were instructed to maintain a minimum compliance level (80%) on average to warrant eligibility for monetary remuneration. Other study details not included here can be found in [13] published shortly after the study started. Data enrollment and collection began in January 2018 and continued through April 2019 with a median data collection time of 336 days for participants.

A. Infrastructure - Wearable Data

For each sensing sub-stream (phone agent, wearable, beacons, social media), a specific software stack was developed for the purpose of gathering data through the sensors. In particular, we focus on the software stack associated with the wearable for the participants, namely the Garmin vivoSmart 3. Notably, we selected the Garmin vivoSmart 3 as it was the only wearable at the time that offered: (1) a reasonable battery duration between full charges (4 to 5 days); (2) a reasonable charge time (charging during showering was typically sufficient); (3) amenability to capturing sleep due to battery duration and built-in capabilities; and (4) the ability to compute heart rate variability (HRV) through beat-to-beat intervals (BBI). As BBI streaming does require a separate license and is not available through the normal Garmin Health API, we only explore data consistency from the Health API callbacks as might be observed with a Fitbit Charge HR [3] or other similar consumer-grade device, which do not include BBI data.

During enrollment, participants needed to install the Garmin Connect app on their phones. Once users had successfully installed the Garmin Connect app and paired the wearable with their smartphone, users were directed to share data access to the study through our study portal website with a link to provide us access to their Garmin cloud data as shown in Fig. 1.

While the procedure was relatively straightforward, there were several issues that emerged. First, users with multiple Garmin Connect accounts could sometimes login and authorize the account that was not linked to their app. Second, users would confuse which account to login via the Garmin Connect authorization, trying to login with their GMail account and password. Third, as many of the users were enrolled remotely, several did not follow all instructions and simply skipped over the authorization step. These resulted in syncs that did not generate callbacks to the backend. For these cases, researchers queried the HealthAPI to recover past data.

Although the vivoSmart 3 itself could store data locally for several days before it needed to sync with the phone, users sometimes would close the Garmin Connect app to save power and / or data. As a result, during our study, we sent emails every Monday to participants that had not synced data for four days to prevent this issue. Typically, there were 120 to 150 participants (15%) who needed a synchronization reminder with those users tending often to repeat.

Data was stored in a PostgreSQL database , with the entirety

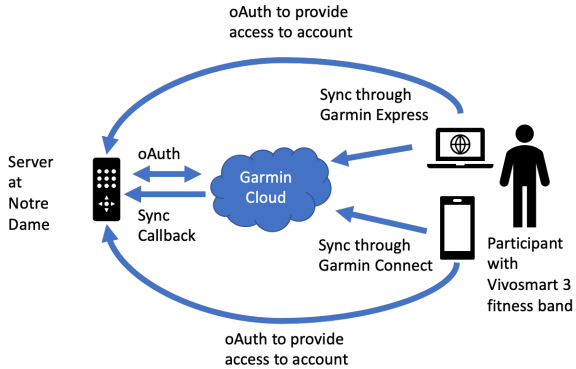


Fig. 1: Sync and authorization procedure to receive data from the smartwatch through the cloud.

TABLE I: Statistics of unique samples of data and number of days with samples for participants in the study by type of data.

Type of Data	Participants		Participants W/ Data (# of samples)						Participants W/ Data (# of days W/ samples)			
	W/ Data	W/o Data	Total	25th	Median	75th	Max	Std	Median	Mean	Max	Std
Activities (I)	543 - 71.7%	214	40,182	2	12	87	1165	137	9	46	362	69
Activities (M)	202 - 26.7%	555	3,434	2	4	15	194	32	4	14	165	28
BodyComps	509 - 67.2%	248	13,743	1	3	24	645	61	3	23	281	46
Dailies	724 - 95.6%	33	207,064	263	319	358	365	96	319	286	365	96
Epochs of Activity	724 - 95.6%	33	19,070,884	23,477	29,521	33,399	35,181	9,285	319	286	365	96
Sleep	711 - 93.9%	46	342,702	343	525	642	1398	233	263	232	362	98
UserMetrics	710 - 93.8%	47	73,130	47	92	157	344	70	92	103	344	70

of the dataset containing approximately 20 million rows taking up 16GB of disk space. In Table I we present the statistics of the summaries, epochs of activity, and user characteristics gathered. A breakdown of what each type of data is follows.

B. Types of Data Collected

For the wearable, the Garmin Health API offered three main types of data: Epochs (heart rate, stress, activity), Summaries (activity, daily, sleep), and User Characteristics.

1) Activities: The Garmin Health API provides inferred user activity summaries as well as manually entered activities. Inferred user activities overlap with motion coefficients submitted in epochs and with daily summaries of activity. This data type is an attempt from the cloud to provide high level data about activities such as statistics on speed, distance traveled, heart rate, and intensity. These activities can be updated by participants using the Garmin Connect app. Manually entered activities are the ones that are created by the user and were not detected by the smartwatch, i.e., the cloud could not infer there was an activity from the data submitted. This feature was only utilized by 202 participants. However, we do not have ground truth that would let us estimate how many samples of activities we should have collected.

2) Body Composition: This refers to a summary containing body weight, muscle mass, bone mass, body water percentage, and body mass index. This data is not collected by the smartwatch automatically. It can be collected in three ways: entering it through the Garmin Connect app; through the MyFitnessPal app; or through a Garmin Index body composition scale. This last option was used by 52 participants a total of 5098 times.

3) Daily Summaries: These data contain summary statistics on daily stress, calories, distance, heart rate, and steps.

4) Epochs: Heart rate and stress score epochs provide one value for each sample and one offset over an initial timestamp to determine to which epoch the sample belongs. When there is data present, each heart rate sample represents 15 seconds in time, while each stress sample represents 3 minutes in time. Gaps in the data are represented by a value of -1 and the length of the gap may exceed 15 seconds. We gathered approximately 1 billion samples of heart rate and 178 million samples of Garmin's stress score throughout the study.

Epochs of activity break down the physical activity of participants into 15 minute epochs. However, if a participant has carried on multiple activities such as walking, sitting, and running during those 15 minutes, one sample for each will be received at the backend, each one having a duration of 15 minutes with their reported active time adding up to 15 minutes. Each activity sample will report: an active time; active Kilocalories burnt; distance traveled in meters; Metabolic Equivalent of Task [14]; mean and max motion intensity; steps; the result of classifying the activity as one of walking, sedentary, running, sleep, generic, or unmonitored; and a classification of motion intensity into sedentary, active, or highly active.

5) Sleep: Sleep summaries contain the duration of the sleep, a distribution of the periods of light sleep, deep sleep, REM sleep, awake time, and unmeasured time during the sample of sleep. Unmeasured time may or may not correspond with off-wrist time according to documentation.

6) User Metrics: Garmin infers user metrics from the activities performed by participants while wearing the Vivosmart 3. This includes an estimate of maximal metabolic rate (VO2max) and a "fitness age". To obtain a fitness age, Garmin compares internal fitness metrics with similar participants by age and gender. An age of X for a participant of gender Y, means that the fitness level is comparable to the average user of age X and gender Y.

IV. Results

In our study we defined wearable compliance to be an estimate of the percentage of time that participants were wearing the watch. Therefore, we study compliance and use this term interchangeably with wearing time.

A. Computing Wearable Compliance

As reported in [2], sometimes a watch may not have a sample of heart rate for a specific period of time because of the lack of accuracy of the watch to determine one. As a consequence, a lack of an HR sample is not sufficient to determine that the user is not wearing the watch. In our study, however, the risk of falsely determining that the participant is wearing the watch is low with respect to other smartwatches like the original Apple Watch, because the vivoSmart 3 will go into a deep sleep mode if it is not on a wearer's wrist, mitigating one of the concerns presented by [2].

We defined wearable compliance in a given day as the ratio of the number of 30-minute windows in the day that had at least one data point, to the total number of windows in a day, regardless of the type of sample that the window contained Given that other measures aside from heart rate were being

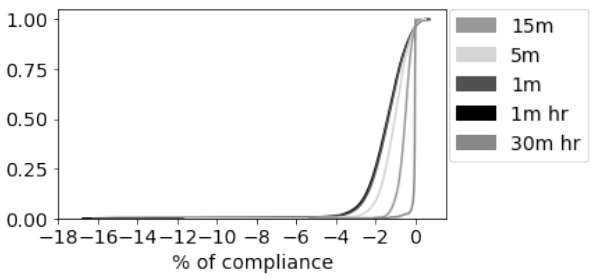


Fig. 2: CDF of the mean difference of compliance between using 15 min, 5 min, 1 min, 30 min heart rate only, or 1 min heart rate only time windows, versus 30 min all data types time windows

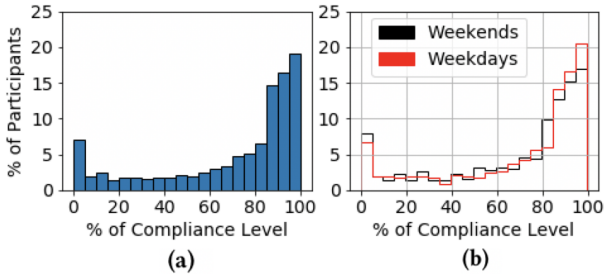


Fig. 3: (a) Compliance distribution in the study. (b) Compliance distribution in weekends vs weekdays (right)

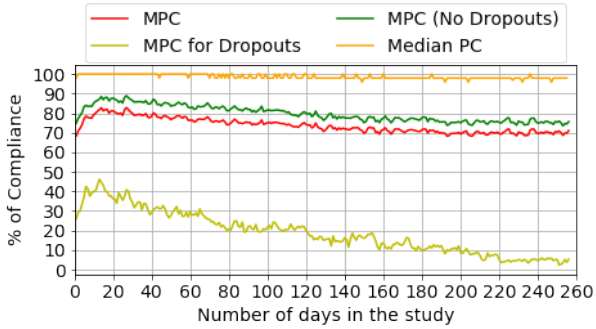


Fig. 4: Mean Participant Compliance (MPC) and Median Participant Compliance by participants' day in the study

sampled at a slower and unpredictable rate, we intuitively chose a large window to minimize the impact of missing heart rate data on participants' compliance while the watch was being worn, even if it meant overestimating the time that the participant had been wearing the watch.

For the purpose of this work, we computed compliance in shorter time windows of 1 minute, 5 minutes, 10 minutes, and 15 minutes to determine if we overestimated compliance by a wide margin. Additionally, we computed compliance using only heart rate data to compare the results. For the vast majority of participants, the use of a shorter time window to calculate compliance resulted in a difference of less than 4%, as presented in Fig. 2. Given the lack of ground truth regarding watch takeoff events, we could not determine if missing data when calculating compliance with a higher resolution meant that the watch was not being worn at the time. Despite this, we looked at the greatest differences in mean compliance, which were between 1 minute and 30 minute time windows: M=71.82% SD=24.91 using 1 minute windows, while M=73.33% SD=27.63 when using 30 minute windows. We consider this overestimation of 1.51% to be a reasonable trade-off given that calculating compliance with a longer time window has the benefit of having less data to store per user and in our case it reduced the computation time.

We arrived to similar results when calculating compliance using only heart rate data, instead of all types, and as shown in Fig. 2. Compliance using only heart rate samples was M=71.77% SD=27.39% when using 1 minute time windows and M=73.28%, SD=27.64% when using 30 minute windows. There is a negligible difference of 0.05% between using all types of data and using only heart rate. This means that it is unlikely that the watch captures any sample of any data type if it is not capturing HR at the same time.

B. Compliance Results

Our compliance distribution in Fig. 3a shows that most of our participants were compliant, with more than 50% of the users having worn the Garmin watch for more than 80% of the time. Our average compliance was 73.5% and our median was 97.9%. Averaging across the average of each participant, thus ignoring differences in how many days the participants were in the study, results in a mean compliance of 70.0%, and a median of 85.0%. If we remove participants that never submitted any data (i.e., compliance = 0), then the mean would be 73.3% and the median would be 85.9%.

1) Gaps in the Data: A gap in the data is a period of one or more windows of time where we did not receive a sample from the wearable of a participant. Short gaps in the data are bound to happen because, at the very least, the vivoSmart 3 cannot be charged while it is being worn. The study of the gaps in our data revealed that 91.1% of the gaps lasted less than a day, 50.1% less than 3 hours, and the average number of gaps per day was 1.14. Nevertheless, gaps longer than 24 hours account for 77.2% of all data missed. If we focus on gaps in participants that did not drop out of the study we find that 70.8% of their missing data happened in periods of time longer than 24 hours and 45.1% in periods longer than a week. The latter affected 311 participants or 47.4% of the participants in the study. Even though we can see in Fig. 6a that some participants had these gaps multiple times, we think that breakages are behind the majority of the cases where these gaps occur.

2) Compliance versus duration in study: The minimum amount of time that a participant that did not drop out of the study was in the study was 257 days. Therefore, if we look at compliance for participants that have been in the study for at least 257 days as in Fig. 4, we see that their compliance from the moment they joined the study grows in the beginning and then decreases the longer the participants stay in the study before settling around 70%. Likewise, if we remove dropouts from consideration, then the compliance follows a similar

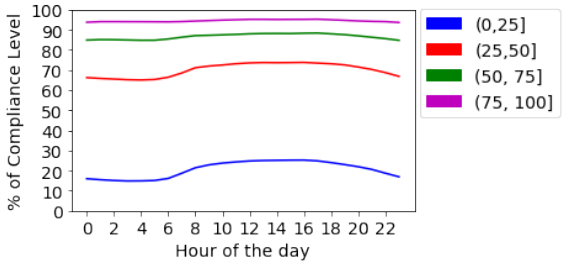
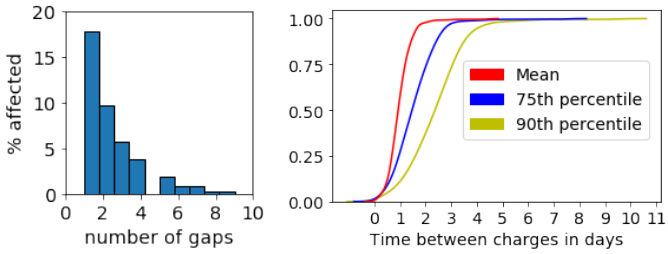
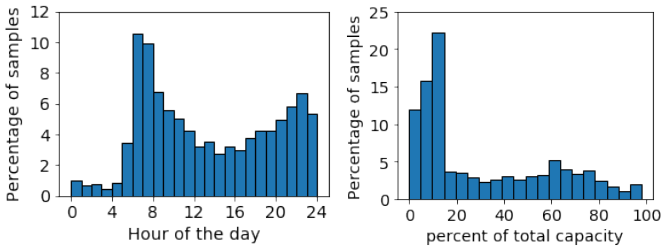


Fig. 5: Compliance by hour of day for participants grouped by quantiles in the compliance distribution: 0 to 25%, 25% to 50%, 50% to 75%, and 75% to 100%



(a) (b)

Fig. 6: (a) Percentage of users affected by gaps longer than a week by number of times that they were affected (b) CDF of time between charging events: mean time per user, 75th percentile per user, 90th percentile per user.



(a) (b)

Fig. 7: (a) Charging events started per hour of the day. (b) Battery level before charging

trend although it settles close to 75%. We can see that the median participant compliance remained stable and close to 100% throughout the entire study.

C. Temporal Patterns of Wearing Behavior

We studied compliance within the week and compliance within the day. There was a slight difference in the compliance between weekdays and weekends as noted in Fig. 3b. While [2] found that the difference in compliance between weekdays and weekends was 10.7%, in our data there was a mean difference of only 2.4% (Weekdays: M=69.9%, SD=41.4%; $t(268559)=13.67$, $p<0.001$).

1) Compliance by hour of the day: We can see in Fig. 5 that there is usually lower compliance overnight than there is throughout the day. The pattern is more obvious in participants with low compliance and is consistent with having some participants taking off their devices at night on occasion. Some participants reported in communications with researchers not wearing the device at night but as Fig. 5 shows, it was not a widespread issue with even the lower quartile seeing no more than 10% difference on average between nighttime and daytime.

D. Charging the Wearable

During month 10 in the study, a Garmin Battery Level stream was added to the iOS phone agent used in the study. This allowed us to collect charging data for 245 iOS participants. By analyzing the cumulative distribution function of the charging events from the study, we found that participants did not fully charge and discharge the watch on most occasions. The majority of the participants, on average, charged their device more frequently than once every 4 to 5 days as evidenced in Fig. 6b. Intuitively, we would expect a priori a correlation between the time that the device was not being charged and how compliant a participant was. For the device to capture data it needs to be on the wrist and the vivoSmart 3 cannot be charged while it is being worn. However, a Pearson correlation revealed an R squared of 2.89%, $p<0.001$ between the average charging time and the average compliance. This correlation suggests that charging could only account for 2.89% of the variance in compliance. This finding and the fact that most data got missed in long gaps at a time, point to other behaviors or issues outside of the participants' control such as breakages, explaining the variance in compliance.

E. Accuracy of the Wearable

During enrollment, 136 participants volunteered to wear a Zephyr Bioharness 3 during 30-60 minutes. From those participants, 31 had overlapping heart rate (HR) samples from the wearable, covering 13.7 minutes on average per participant. The vivoSmart's sampling frequency was of $1/15$Hz. The bioharness' sampling frequency was 1Hz. To compare the measurements, we used two windowing methods. First, we averaged the bioharness' samples across 15s fixed windows such that it would match the rate of the watch. Second, we used 5-minute sliding windows, sliding 1 minute at a time, because that is the way that the data was ultimately used for data analysis in the study.

In the case of 15s windows, the results show a fair ICC (0.44; 95% CI 0.27 to 0.56), an MAE of 10.4 bpm and a MAPE of 13.1%. Although the agreement is not high, it matches the findings in [15]. In the case of 5-minute windows, the results show a similar ICC (0.43; 95% CI 0.27 to 0.55), although a slightly better MAE of 8.4 bpm and MAPE of 10.5%.

F. Study maintenance

Because of the scale and duration of our study we expected devices to break or be lost. We share our experience and break down the issues that forced us to replace 325 devices or chargers:

- 209 were replaced due to the watch strap breaking.

- 66 were reported to not hold charge, not charge despite being connected to the charger, not sync data, report unusually high or low numbers of steps or floors climbed, or inability to connect to the phone.
- 17 were replaced due to the loss or problems with the charger.
- 10 were replaced due to having issues with the screen such as the display not working correctly or the screen cracking.
- 22 were replaced due to losing the wearable.
- 1 participant reported an allergic reaction to the nickel in the buckle.

We believe breakages, more than user charging behaviors, were responsible for the long gaps of missing data that we found in our time series.

V. Conclusion

We have studied the wearing behavior of participants using our calculation of compliance as an estimate of the wearing time of participants. Participants did not show a large difference in compliance by hour of the day as in [2]. Participants wore their smartwatches 73.5% of the time in the study, more time than in our previous study involving students [3]. The majority of the time that participants were not wearing the watch was found to be in contiguous stretches of several days and weeks at a time, consistent with a likely cause being the 325 breakages that we suffered during the study. The majority of within day gaps were shorter than an hour. Analysis of charging behavior found that only half of the time users waited for battery capacity to reach 15% or less before charging, and in the other half capacity was distributed uniformly. Compliance during the week showed a slight difference of 2% in weekends vs weekdays. Finally, compliance for participants that did not drop from the study decreased steadily from the peak of 90% that happened during the first month before stabilizing at over 70% close to day 190. Although sometimes inaccurate in their measurements as we found, inexpensive smartwatches remain convenient for capturing physical data in the real world without incurring in a high time demand from busy workers allowing researchers to get reasonable quality longitudinal data.

Our findings can help guide future study design in the estimation of participants, budget, and running time needed to achieve a certain level of data collection. Our experience indicates that fitness bands can achieve high percentages of data collection and reducing missing data further would likely involve a greater logistic effort to replace devices faster. Future work will focus on providing a multi-variable model based on the previously mentioned traits and factors included in the ground truth instruments of the study to predict compliance and attrition at the beginning of a study with the goal of maximizing data collection in future studies.

VI. Acknowledgements

This research is 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] I. D. Corporation, "IDC Reports Strong Growth in the Worldwide Wearables Market..." Accessed: April 2019.

[2] H. Jeong, H. Kim, R. Kim, U. Lee, and Y. Jeong, "Smartwatch Wearing Behavior Analysis: A Longitudinal Study," *Proc. of the ACM on IMWUT*, vol. 1, no. 3, pp. 1–31, Sep. 2017.

[3] 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," ser. ISWC '16. New York, NY, USA: ACM, 2016, pp. 28–35.

[4] D. Harrison, P. Marshall, N. Bianchi-Berthouze, and J. Bird, "Tracking Physical Activity: Problems Related to Running Longitudinal Studies with Commercial Devices," *Ubicomp 2014*, Sep. 2014.

[5] R. Rawassizadeh, E. Momeni, C. Dobbins, P. Mirza-Babaei, and R. Rahnamoun, "Lesson Learned from Collecting Quantified Self Information via Mobile and Wearable Devices," *Journal of Sensor and Actuator Networks*, vol. 4, no. 4, pp. 315–335, Nov. 2015.

[6] 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. ACM, 2014, pp. 3–14.

[7] M. De Choudhury and S. Counts, "Understanding affect in the workplace via social media," ser. CSCW '13. ACM, 2013, pp. 303–316.

[8] V. W.-S. Tseng, S. Abdullah, J. Costa, and T. Choudhury, "Alertnessscanner: What do your pupils tell about your alertness," ser. MobileHCI '18. New York, NY, USA: ACM, 2018, pp. 41:1–41:11.

[9] S. Liu, Y. Jiang, and A. Striegel, "Face-to-face proximity estimation using bluetooth on smartphones," *IEEE Trans. on Mobile Computing*, vol. 13, no. 4, pp. 811–823, April 2014.

[10] G. M. Harari, S. R. Muller, V. Mishra, R. Wang, A. T. Campbell, P. J. Rentfrow, and S. D. Gosling, "An Evaluation of Students' Interest in and Compliance With Self-Tracking Methods," *Social Psychological and Personality Science*, vol. 8, no. 5, pp. 479–492, Jul. 2017.

[11] L. Faust, R. Purta, D. Hachen, A. Striegel, C. Poellabauer, O. Lizardo, and N. V. Chawla, "Exploring Compliance: Observations from a Large Scale Fitbit Study," ser. SocialSens'17. New York, NY, USA: ACM, 2017, pp. 55–60.

[12] O. Crowley, L. Pugliese, and S. Kachnowski, "The Impact of Wearable Device Enabled Health Initiative on Physical Activity and Sleep," *Cureus*, vol. 8, no. 10, 2016.

[13] S. M. Mattingly, J. M. Gregg, and e. al., "The Tesserae Project: Large-Scale, Longitudinal, In Situ, Multimodal Sensing of Information Workers," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '19. ACM, 2019, pp. CS11:1–CS11:8.

[14] C. for Desease Control US Government, "General Physical Activities Defined by Level of Intensity," *Online*, p. 5, Accessed: April 2019.

[15] A. M. Müller, N. X. Wang, and e. al., "Heart rate measures from wrist-worn activity trackers in a laboratory and free-living setting: Validation study," *JMIR Mhealth Uhealth*, vol. 7, no. 10, p. e14120, Oct 2019.