This full text paper was peer-reviewed at the direction of IEEE Instrumentation and Measurement Society prior to the acceptance and publication.

Measuring the Fitness of Fitness Trackers

Chelsea G. Bender, Jason C. Hoffstot
Brian T. Combs and Sara Hooshangi
Integrated Information, Science, and Technology
The George Washington University
Washington, DC, USA
Email: shoosh@gwu.edu

Justin Cappos
Computer Science and Engineering
New York University
New York, NY, USA
Email: jcappos@nyu.edu

***Abstract*—Data collected by fitness trackers could play an important role in improving the health and well-being of the individuals who wear them. Many insurance companies even offer monetary rewards to participants who meet certain steps or calorie goals. However, in order for it to be useful, the collected data must be accurate and also reflect real-world performance. While previous studies have compared step counts data in controlled laboratory environments for limited periods of time, few studies have been done to measure performance over longer periods of time, while the subject does real-world activities. There are also few direct comparisons of a range of health indicators on different fitness tracking devices. In this study, we compared step counts, calories burned, and miles travelled data collected by three pairs of fitness trackers over a 14-day time period in free-living conditions. Our work indicates that the number of steps reported by different devices worn simultaneously could vary as much as 26%. At the same time, the variations seen in distance travelled, based on the step count, followed the same trends. Little correlation was found between the number of calories burned and the variations seen in the step count across multiple devices. Our results demonstrate that the reporting of health indicators, such as calories burned and miles travelled, are heavily dependent on the device itself, as well as the manufacturer's proprietary algorithm to calculate or infer such data. As a result, it is difficult to use such measurements as an accurate predictor of health outcomes, or to develop a consistent criteria to rate the performance of such devices in head-to-head comparisons.**

Keywords—Fitness Trackers; Accuracy; Physical Activity; Free-living Conditions

I. Introduction

The past several years have seen an exponential growth in the market for personal wearable devices, with estimated sales of up to 126 million units anticipated by the end of 2019 [1]. Fitness tracking devices lead sales in this market, and continue to gain popularity as the correlation between an active lifestyle and the prevention of chronic diseases is demonstrated by research [2], [3]. These trackers give their users the ability to monitor and track key health markers, thus encouraging them to continue their healthy efforts.

As manufacturers try to improve the accuracy of these health measurements by adding functionality and introducing new devices into the market place at a rapid pace, the number of ways that this collected and stored data can be used also increases. Individuals can use data on their average daily/weekly physical activity to monitor their own health, or to identify key markers to report to their health providers. Public health researchers could use such data in aggregated form, in large-scale studies to monitor health related outcomes for different segments of the population. And, on a larger scale, programs sponsored by insurance companies can promote healthier lifestyles by offering incentivizing discounts on life and health insurance products based on the physical activity levels of consumers.

Such programs, however, rely on the ability of these devices to reliably generate accurate data. Data accuracy ultimately depends on two factors: the quality of the sensors embedded in the device, and the algorithm used to interpret the raw data. To this end, there has been a surge in the number of research studies testing the accuracy of wearable fitness devices as compared to research-grade accelerometers and multi-sensor devices [4]–[11]. Most of these studies have focused on a cross-sectional comparison of consumer-based products to research-grade gold standards only in a laboratory or a controlled real-world environment [4]–[7], [12]. Conducting experiments without the prescribed restrictions of a laboratory (i.e. under a free-living condition) is significantly more challenging, as the variations in speed, direction and intensity of physical activities are larger. This may be why only a few studies have measured the accuracy of trackers in free-living conditions [8]–[10] and most free-living studies have been short in duration, typically in the range of one or two days. Furthermore, the integrity of these results could also be compromised if the subjects under study (who are often volunteers) do not follow the experiment protocols.

In this work, we set out to compare parameters and experimental settings that have not been explored in previous work. We start by looking at other health indicators measured by these devices, such as calories burned or distance travelled. We designed a series of experiments to compare these along with the more commonly studied measure of step counts. While the step count provides a general sense of movement and physical activity, calories burned and the number of miles travelled could be better indicators of an individual's energy expenditure and, hence his/her physical fitness level. If the fitness trackers are to become an integral part of our health-

978-1-5090-3202-0/17/$31.00 ©2017 IEEE

monitoring regimen, the accuracy of all data must be validated.

Two other factors that set our research apart from most previous efforts is that we ran our experiments in free-living conditions for a longer time period than all other previous studies. All three experiments ran for 14 days, and in each, the subjects wore two devices on the same wrist as they went about performing daily life activities. The exact position of the device on the wrist was switched every few days (i.e. the device worn closer to the wrist on one day was worn further from the wrist on the same arm on a different day) in order to eliminate the dependency of the result on the exact location on the wrist. The devices were removed when subjects went to sleep. In the first study, two identical Fitbit Flex trackers *(Fitbit, Inc., San Francisco, CA, USA)* were examined. In the second experiment a Fitbit Charge HR was compared to Garmin vívoactive *(Garmin Ltd., Olathe, Kansas, USA)*, and for the third study an Apple Watch *(Apple Inc., Cupertino, California, USA)* was tested against the Fitbit Flex fitness tracker. In addition, since the subjects who participated in these experiments were part of the research team, they were able to follow appropriate protocols.

The results of our study suggest that measurements for all three data categories examined could vary significantly when compared side by side. While the variations for step count and miles travelled followed the same trends, there was no apparent correlation between variations in calories count and that of the other categories. Therefore, it is important to take such variations into account when implementing programs that could rely on the accuracy of a variety of fitness devices

The rest of this paper is organized as follows. Section II gives an overview of some of the research in validating and comparing data tracked by fitness devices. Sections III and IV describe the experiment design, methodology and the analysis of the results. Section V discusses future work, and Section VI concludes the paper with some comments on what was learned from our study.

II. Literature Review

A wide range of fitness activity monitors has flooded the market over the past five years, providing researchers in exercise science, nutrition, and sport medicine with new measurement tools. But, before such tools can be incorporated into research, the accuracy of the data must be validated. Several studies have examined the accuracy of as many as ten fitness monitors simultaneously, by comparing the number of steps reported by these consumer-based products [4]–[11].

In one study [10], ten consumer-level wearable fitness devices were examined, both in the laboratory and in free-living conditions, and results were compared to a research-grade pedometer. This study focused solely on the comparison of step counts across the ten devices. Under laboratory conditions, the participants walked on a treadmill for 30 minutes wearing all ten consumer-based devices and two research-grade devices on two different days. Under free-living conditions, the participants wore all consumer devices and only one research-grade device (ActivPAL) for seven and half hours on a single day. They concluded that seven of the ten devices showed similar output when counting steps, and five showed a relatively close output compared to the research-grade device.

In a different study [9], seven consumer-level wearable fitness devices were compared to two research-grade devices during a 48-hour period timeframe in free-living conditions. This study measured step count, and other parameters such as the total daily energy expenditure (TDEE). The team found that the measured steps for all consumer-level devices had a strong correlation with those of the research-grade devices, while all the consumer devices greatly underestimated TDEE compared to the research-grade devices.

III. Data Collection

A. Experimental Setup

Since the above studies confirmed that the step count of consumer-base devices are comparable to measurements obtained from research-grade devices, the focus of this study was to compare performance and data accuracy across a series of consumer-based products. We chose fitness trackers that had a large share of sales in the market at the time of our experiment [13]: Fitbit Flex, Fitbit Charge HR, Garmin vívoactive, and Apple Watch. Table 1 showcases some of the common data collected by these fitness trackers. The combination of motion and direction sensed by an onboard tri-axial accelerometer and a gyroscope are used to calculate the number of steps and flights of stairs taken. The number of calories burned can be inferred from this information using an internal algorithm that might vary between manufacturers.

TABLE I
FITNESS DEVICES USED IN THIS STUDY

Make-Model	**Device Description**
Fitbit Flex	Tracks steps, distance, calories burned, and active minutes. Also, how long and well the user sleeps.
Fitbit Charge HR	Provides continuous, automatic, wrist-based heart rate and simplified heart rate zones. Tracks workouts, heart rate, distance, calories burned, floors climbed, active minutes and steps. Monitors sleep automatically and has an alarm.
Garmin vívoactive	Built-in sports apps, including GPS-enabled running, biking and golfing options, and swimming and activity tracking.
Apple Watch	A smartwatch that claims to monitor every move, not just walking or running. Apple Watch collects data when the user stands, sits, and exercises through GPS and calculates heart rate.

B. Data Collection Methodology

Three independent experiments were conducted during this study. A different member of the research team wore two

devices side by side on the same wrist for 14 days. The research team included two males and a female. The devices were worn during the day, from the time that the subjects woke up until the time that they went to bed for an average daily wear time of 16 hours. Before the start of the experiments, each subject entered the required physical information, such as gender, height, weight, and age, via the manufacturer's mobile application or website.. At the completion of each day, the subjects would synchronize the devices with either the mobile application or the cloud service application associated with each manufacturer in order to submit their daily activities.

In order to eliminate location-based dependencies, the subjects switched the order of the devices on the wrist every few days. In the first experiment, a male member of the research team wore two identical Fitbit devices (Fitbit Flex) side by side on the dominant wrist (right) for two weeks. In the second experiment, a female member of the research team wore a Fitbit Charge HR and a Garmin vívoactive tracker on her non-dominant hand (left) for two weeks. In the third experiment, a male member of the research team wore the devices on his non-dominant hand (left) for two week. Table 2 summarizes the devices used in these three experiments.

TABLE II
EXPERIMENT DESIGN

	Devices Under Study
Experiment 1	Fitbit Flex & Fitbit Flex (Fitbit 1- Fitbit 2)
Experiment 2	Fitbit Charge HR & Garmin vívoactive
Experiment 3	Fitbit Flex & Apple Watch

IV. Results and analysis

A. Experiment 1: Comparing devices of same model

This experiment was performed as an inter-device study, and as a baseline to assess the reliability of Fitbit Flex, which is a wristband tracker. The results are depicted in Figure 1, where the relative differences (Fitbit 1 less Fitbit 2) in number of steps taken, calories burned, and miles travelled are plotted against time. The data is sorted in ascending order of step count variations, with the highest difference in step count being 7%. It is worth noting that, throughout the experiment, Fitbit 1 always showed a higher step count than Fitbit 2. We did not observe a strong correlation with the daily relative differences and the absolute number of steps taken (i.e. higher number of steps per day did not translate into a higher relative difference).

Previous studies have all indicated that Fitbit has good alignment with research-grade devices and have confirmed the accuracy of step counts for this manufacturer [8], [10]. Our two-week experiment in free-living conditions also shows that the two identical devices are consistent in counting the number of steps, with little variation in either the number of steps or reported distance travelled for the two devices.

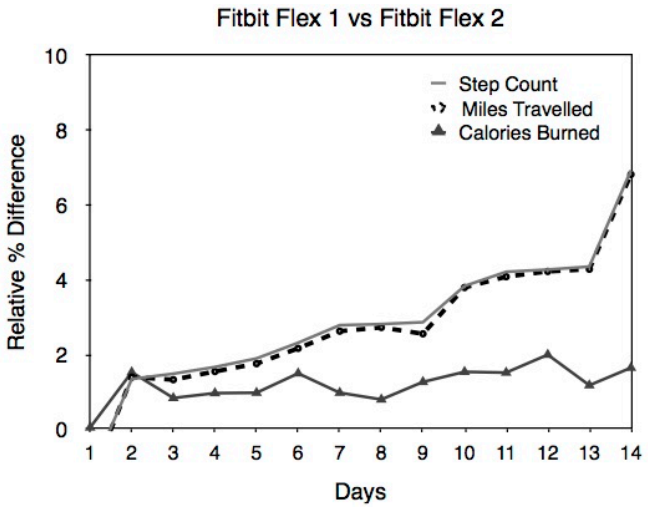

Fig. 1. Comparison of two Fitbit Flex devices over a two-week period. The daily relative percentage differences (Fitbit 1 minus Fitbit 2 divided by the mean of the two) in the number of steps taken, total calories burned, and miles travelled as reported by the device are shown over 14 days. The data is sorted by the relative difference in step count.

One interesting observation shows no correlation between the number of steps taken each day, and the reported number of calories. Even for the days where we see the highest variations in the step count, the deviations in the reported calories were only around 1.6%. We suspect that this may be related to the non-linear model used to calculate calories burned. The Fitbit calorie count resets each night at midnight and begins counting immediately thereafter. As a result, without even getting out of bed, each morning a basal metabolic rate of 700-1000 is registered for an individual based on gender, age, height, and weight. This rate could account for about half of the wearer's daily calorie consumption. As a result, small variations in the step count do not translate to a significant change on the reported calories burned.

B. Experiment 2: Comparing different brands

This experiment was done to compare two smartwatches from different brands. We examined Fitbit Charge HR and Garmin vívoactive. The results of our two-week experiment are depicted in Figure 2, where the relative differences (Fitbit Charge HR less Garmin) in the number of steps, calories burned, and miles travelled are plotted against time. The data is sorted in ascending order of difference in step counts, with the highest difference peaking at 34%. We noticed that, throughout the two weeks of this experiment, the Fitbit device consistently showed a higher step count than Garmin, regardless of the location of the devices on the wrist (i.e. it did not matter which device was worn closer to the wrist).

Similar to Experiment 1, we did not observe any strong correlation between the daily relative differences and the absolute number of steps taken. The number of miles travelled showed the same trend, but with an overall lower variation, up

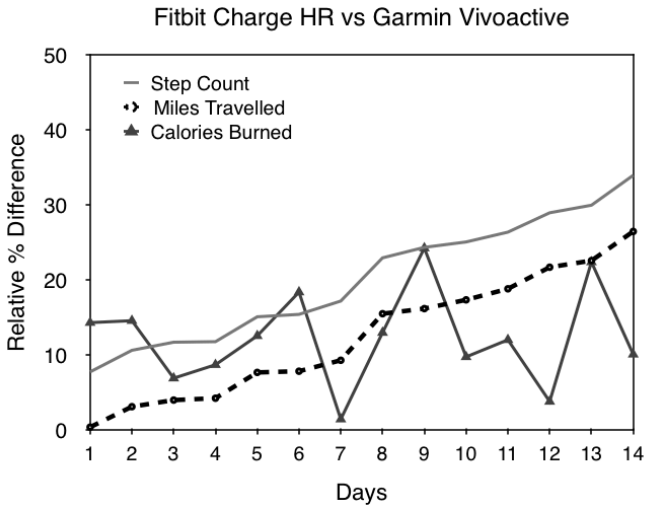

Fig. 2. Comparison of Fitbit Charge HR and Garmin vívoactive over a two-week period. The daily relative percentage differences (Fitbit minus Garmin divided by the mean of the two) in the number of steps taken, total calories burned, and miles travelled as reported by the device are shown over 14 days. The data is sorted by the relative difference in step count.

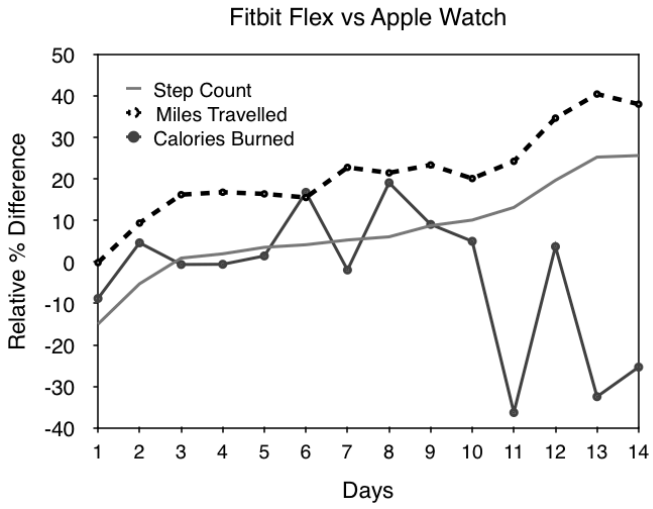

Fig. 3. Comparison of Apple Watch and Fitbit Flex over a two-week period. The daily relative percentage differences (Apple minus Fitbit divided by the mean of the two) in the number of steps taken, total calories burned, and miles travelled as reported by the device are shown over 14 days.The data is sorted by the relative difference in step count.

to a maximum of 26%. The surprising result was the lack of correlation between the number of steps taken and the amount of calories burned. Over the two-week period we did not observe any consistent pattern between the variation in calorie count and step count. We conclude that, in this experiment, a cross-device comparison of the measured calories might not result in an accurate prediction or indication of health levels.

C. Experiment 3: Measuring smartwatch fidelity

This experiment was performed to understand how a multi-use device, like the Apple Watch performs. As a baseline, we compared it against the Fitbit Flex. The result of our two-week long experiment is depicted in Figure 3, where the relative differences (Apple less Fitbit) in the number of steps, calories burned, and miles travelled are plotted against time. The data is sorted in ascending order by the difference in step counts, with the highest and lowest difference registering at 26% and -15%, respectively. Throughout most of the experiment (with the exception of two days), Apple Watch showed a higher step count than the Fitbit, regardless of the location of the device on the wrist, much like what was observed in Experiments 1 and 2.

In this experiment, the variation in the number of miles traveled is higher than the reported step count. However, similar to Experiment 2, there is no significant correlation between calorie variations and either miles travelled or steps taken. This experiment further affirms that these variations are seen across different fitness trackers. As a result, such measurements need to be better understood if reliable cross-comparison studies are to be done, or if health providers and insurance companies choose to rely on such data to provide services or rewards to their consumers.

D. Analysis

Our first experiment confirms the reliability of Fitbit Flex in reporting the daily number of steps taken by an individual. Our result correlates well with a previous study that reported on the inter-device reliability of a different model of Fitbit [11]. Our two-week experiment further affirms the reliability of Fitbit over a longer period and outside of a controlled lab environment. We infer that individuals can reliably compare their daily physical activity with peers who own the same Fitbit model. While a further study is needed to confirm the reliability of Fitbit devices across different models, we expect this inference would be supported.

Our results also indicate that the calculated distance travelled has a high correlation with the number of steps taken. This is expected, as Fitbit uses a simple linear model to calculate the travelled distance by multiplying walking step count and the stride length. On the other hand, the deviation in calorie counts does not correlate well with the number of steps taken, where a 7% variation in the step count corresponds to a 1.6% variation in the calorie count. We suspect that this may be related to the algorithm that is used to calculate calories. Fitbit estimates a Basal Metabolic Rate (BMR) for an individual, based on data entered during account setup such as gender, age, height, and weight. The BMR usually accounts for at least half of the reported daily calories, but the exact manner by which the step count is integrated into this equation is not disclosed.

Experiment 2 compares the results of two very different fitness trackers with similar output measures. The variations in the step count is much higher than what was seen in

Experiment 1. Some of the observed variations could be related to the actual sensors used in each device. We are not aware of the specification of the tri-axial sensors used in each device, but we suspect that they are manufactured or calibrated differently. Previous studies and our observations indicate that wrist-worn fitness trackers show higher levels of variations when compared to the more conventional hip-worn trackers used in laboratory-based experiments. This can be attributed to the extrinsic noise introduced as a result of using hand movement as a measure of step count.

Another interesting result of this experiment is the 8% difference in the variation levels of step count and miles travelled between these two devices. We speculate that Garmin GPS functionality might play a role in calculating the underestimated value of the reported distance travelled by vívoactive. However, we were not able to find any information about Garmin's methodology on their website or in the owner's manual. The only qualitative data that we found was several forum postings in which users reported that their devices had underestimated the number of miles travelled, especially when running or engaging in more rigorous physical activity. A further experiment is required to establish a baseline for the accuracy of the reported distance travelled.

While the variations shown in the calculated distances are less than the variations observed in the step counts, they follow the same trend. For example, for days for which there is a higher variation in the step count, we will also see a higher variation in the distance travelled. This is not the case for the calculated calories burned. We speculate that these devices are using different metabolic formulas to calculate the energy expenditure of the user, resulting in very different outputs for the calories burned. Fitbit specification indicates that the heart rate monitor, which in case of Fitbit Charge HR is integrated in the device, is a factor in calculating the calories burned. Unfortunately, our subject reported a high variations in the heart rate reading depending on the tightness of the band around the wrist, which may have affected the calories count as well. A follow-up study to compare the Fitbit Flex with Fitbit Charge HR can shed more light about this dependency.

Experiment 3 compares the Apple Watch with the fitness band Fitbit Flex. Our experiment indicates that the step counts reported by the Apple Watch are higher than the Fitbit Flex. In two laboratory studies [10], [12] Fitbit Flex has shown relatively close agreement with research-grade devices (within a 10% range). As a result, we speculate that Apple Watch might be overestimating the number of steps taken. Unlike the second experiment, we see higher variations in the number of miles reported by these devices. The methodology used by Apple Watch to calculate the distance is not clear to us, but because of its sensitivity, we suspect that other parameters, such as GPS and the movement of the iPhone connected to the watch might also be contributing factors in the calculation of the number of miles travelled. Similar to Experiment 2, we see a rather inconsistent pattern for the number of calories burned. It is worth noting that Apple Watch separates calories into resting and active categories, and the combination of these two measures (based on the level of movement and other queues collected by the watch) determine the total number of calories burned. This might play a role in the large variation seen for this parameter.

V. Future Work

We plan to continue with data collection over various conditions, for example wearing devices on different parts of the body. A further study is also planned to confirm the reliability of different models produced by the same manufacturer. We will also look more closely at the algorithm that different vendors use to calculate health related attributes. Some of the newer devices provide information about heart rate and sleep patterns, and we plan to better examine those as well.

VI. Conclusion

In this paper, we described a series of experiments where several fitness tracking devices, including two models of Fitbit, a Garmin smartwatch, and an Apple Watch, were used to collect data for 14 days. Data on the number of steps taken, distance travelled and calories burned by each subject was collected over this period, and a comparison analysis was performed. Our data analysis shows that step count, miles travelled, and calories burned could vary significantly when devices of different manufacturers are compared side by side. While the variations in the step count and the distance travelled followed the same trends, we saw no correlation between the variations in calories burned and what was observed for the step and distance variations. While it is difficult to give a concrete explanation for these observations without a detailed analysis of the embedded sensors in the devices and the algorithms used to calculate the reported data from the raw sensor data, we see no consistency among fitness trackers in reporting physical activities.

What is clear is that using different fitness trackers in engaging in social experiments, such as company-wide step count, miles travelled or calorie goals competitions, or as incentives to receive reduced health benefits might not provide accurate outcomes or fair comparisons for the participants. Employers should take such variations into account when implementing programs that rely on a variety of fitness devices.

References

[1] "IDC Worldwide Quarterly Wearable Device Tracker. International Data Corporation," 2016. [Online]. Available: https://www.idc.com/tracker/showproductinfo.jsp?prod_id=962

[2] I. Lee, E. J. Shiroma, F. Lobelo, P. Puska, S. N. Blair, and P. T. Katzmarzyk, "Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy," *The Lancet*, no. 9838, pp. 219–229, 2007.

[3] D. E. Warburton, C. W. Nicol, and S. S. Bredin, "Health benefits of physical activity: the evidence," *CMAJ*, vol. 174, no. 6, pp. 801–809, 2006.

[4] J. A. Noah, D. K. Spierer, J. Gu, and S. Bronner, "Comparison of steps and energy expenditure assessment in adults of fitbit tracker and ultra to the actical and indirect calorimetry," *J Med Eng Technol.*, vol. 37, no. 7, pp. 456–462, 2013.

[5] K. L. Dannecker, N. A. Sazonova, E. L. Melanson, E. S. Sazonov, and R. C. Browning, "A comparison of energy expenditure estimation of several physical activity monitors," *Med Sci Sports Exerc.*, vol. 45, no. 11, pp. 2105–2112, 2013.

[6] E. Fortune, V. Lugade, M. Morrow, and K. Kaufman, "Validity of using tri-axial accelerometers to measure human movement –part ii: Step counts at a wide range of gait velocities," *Med Eng Phys.*, vol. 36, no. 6, pp. 659–669, 06 2014.

[7] J. Takacs, C. L. Pollock, J. R. Guenther, M. Bahar, C. Napier, and M. A. Hunt, "Validation of the fitbit one activity monitor device during treadmill walking," *J Med Res.*, vol. 17, no. 5, pp. 496–500, 2014.

[8] M. A. Tully, C. McBride, L. Heron, and R. F. Hunter, "The validation of fibit zip^{TM}physical activity monitor as a measure of free-living physical activity," *BMC Res Notes.*, vol. 7, p. 952, 2014.

[9] T. Ferguson, A. V. Rowlands, T. Olds, and C. Maher, "The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study," *Int J Behav Nutr Phys Act.*, vol. 12, no. 1, p. 42, 2015.

[10] T. J. M. Kooiman, M. L. Dontje, S. R. Sprenger, W. P. Krijnen, C. P. van der Schans, and M. de Groot, "Reliability and validity of ten consumer activity trackers," *BMC Sports Sci, Med Reh.*, vol. 7, no. 1, p. 24, 2015.

[11] M. L. Dontje, M. de Groot, R. R. Lengton, C. P. van der Schans, and W. P. Krijnen, "Measuring steps with the fitbit activity tracker: an inter-device reliability study," *J Med Eng Technol*, vol. 39, no. 5, pp. 286–290, 2015.

[12] C. MA, B. HA, V. KG, and P. MS, "Accuracy of smartphone applications and wearable devices for tracking physical activity data," *JAMA*, vol. 313, no. 6, pp. 625–626, 2015.

[13] "Fitbit retains wearables market lead," 2015. [Online]. Available: "http://www.statista.com/chart/3762/wearable-device-shipments/"