Assessing Bouts of Activity Using Modeled Clinically Validated Physical Activity on Commodity Hardware

Caitlin Barrett, Gregory Dominick, Kyle N. Winfree

Abstract— **Human activity can be measured through identification of bouts of activity. The Freedson cut point method used by ActiGraph has become one very common and well accepted standard for estimating times of continuous moderate to vigorous physical activity (MVPA). However, such methods do not directly apply to other data sources such as the Fitbit Flex, a wrist worn wireless pedometer. In previous research by the authors, a model was presented to improve the estimates of physical activity (PA) level in the Fitbit devices. This paper considers the estimates of activity bouts, building on the modeled PA level from the Fitbit Flex as compared to the results from the ActiGraph GT3X. The purpose of this paper is to compare the "gold standard" ActiGraph to modeled Fitbit Freedson methods and to establish normative values of expected errors in bout detection between the two devices and methods, both of which are proxy methods aimed at measuring actual physical activity levels.**

Here we compare bout identification using three measures, the ActiGraph Freedson method, Fitbit Intensity Score, and the modeled Fitbit Freedson using three different outcomes.

First, we compare a baseline of per subject per day number and duration of bouts from an ActiGraph GT3X to the results found from using the same methods on the Intensity Score reported by Fitbit and the modeled Fitbit Freedson method. Next, we compare the difference in duration of bouts identified in each data source matched according to similar start and end times. Finally, we compare the bouts found from the three methods to bouts identified in a self report diary.

I. Introduction

Coinciding with national estimates that indicate approximately 60% of adults track their weight, diet, or exercise, and 21% report using some form of technology or device to track personal health data [1], consumer-based devices have helped promote the "quantified-self" movement [2]. With 29 million registered users worldwide, Fitbit is the most popular brand of wearable activity tracking monitor sold on the commercial market [3], [4]. In addition to having an integrated tri-axial accelerometer, Fitbit devices wirelessly sync processed data to a user's smart-phone Fitbit application (i.e., the Fitbit app) and personal Fitbit web account (i.e., dashboard), enabling users to continuously monitor and track various attributes of their activity and activity bouts in real-time.

However, these initial studies have relied on ActiGraph accelerometers to quantify MVPA outcomes due to limited evidence regarding the accuracy of Fitbit to provide intensity cut-point estimates that are similar to the validated cut-points used in research. For example, a recent study by Dominick and colleagues reported that the Fitbit Flex significantly underestimated the percent of time participants spent in light PA by 34% per day and significantly overestimated the percent of time spent in sedentary, moderate, and vigorous PA by 26%, 3%, and 3% per day, respectively [5]. Many studies have found these inconsistencies in measurement congruence between Fitbit and research grade accelerometers, potentially limiting the utility of Fitbit within research settings. These inconsistencies could also have significant health implications on Fitbit users if the users believe they are meeting the PA guidelines when in fact, they are not.

A possible solution to the issue of "cut point non-equivalence" is to model validated PA levels from raw VM3 accelerometer activity counts using Fitbit intensity data. Therefore, the primary aim of this study was to compare the results of identifying bouts of activity from three different sources: the ActiGraph GT3X, the Fitbit reported Intensity Score, and the modeled Freedson equivalent from the Fitbit raw measures (Fitbit Freedson). These results were then used to evaluate how modeling may impact overall bout detection in an effort to establish a similarity between the devices which may enable researchers to use the Fitbit device in place of the ActiGraph. Doing so will change the scale at which physical activity studies can be conducted, as the Fitbit does not require researcher contact beyond the initial visit. Given the widespread adoption of Fitbit devices, this may allow researchers to conduct such studies without the financial burden of purchasing a device for every participant.

II. Methods

The study design, procedures, instrumentation, and methods to assess PA have been previously described [5] and are briefly summarized here. Study participants consisted of a convenience sample of 19 volunteers recruited from the University of Delaware including healthy adult men (n = 4) and women (n = 15). All participants owned a Fitbit Flex device and provided written informed consent. Before data collection, study approval was obtained by the University of Delaware Institutional Review Board.

Subjects concurrently wore the ActiGraph and Fitbit devices on their dominant side for seven contiguous days. We were unable to make any comparison of wear on the dominant or non-dominant sides of the body. Following a short break of several days, the subjects were invited to participate in another optional seven days of data collection. Subjects were included in this analysis if both the Fitbit and

C. Barrett is a Computer Science graduate student in the School of Informatics, Computing, and Cyber Systems at Northern Arizona University

G. Dominick is an Assistant Professors in Behavioral Health & Nutrition at the University of Delaware.

K. Winfree is an Assistant Professor in the School of Informatics, Computing, and Cyber Systems at Northern Arizona University and is the corresponding author: kyle.winfree@nau.edu.

978-1-5090-4179-4/17/$31.00 ©2017 IEEE

ActiGraph data sources could be verified to be synchronized in time; 16 of the original 19 met this criterion.

The results from the previous modeling study were used here for bout detection [6].

III. Instruments

The ActiGraph GT3X (ActiGraph, Pensacola, FL) is a small, research-grade tri-axial accelerometer used to provide objective estimates of PA behavior in free-living conditions [5]. The standard method of assessing PA level in adult populations, the Freedson VM3 method, was used here to establish known PA level for each minute of data [7]. This method categorizes PA levels as sedentary, light, moderate, vigorous, and very vigorous activities from accelerometer counts detected within a specified epoch. Intensity levels were defined based on the following standard cut-points: sedentary (< 200), light ($200 - 2690$), moderate ($2691 - 6166$), vigorous ($6167 - 9642$), and very vigorous (> 9643). The vigorous and very vigorous categories were combined for this study [8].

The Fitbit Flex is a small wrist worn pedometer, which provides measures of steps per minute, METs, and a Fitbit proprietary Intensity Score. The Intensity Score follows the same concept as that of the modified Freedson VM3 algorithm, with increasing levels of activity corresponding to sedentary, light activity, moderate, and vigorous levels. Note that the Fitbit Flex does not report a very vigorous category. From the previous research by the authors, a Fitbit Freedson PA measure was also established. Here, the expected PA level reported by the ActiGraph GT3X was modeled using Fitbit raw measures. In that previous research, it was shown that this modeling approach significantly improved the estimates of PA on the Fitbit, to the point where the modeled time in MVPA was no longer significantly different from that of the ActiGraph device. It is important to note that the Fitbit reported nearly twice as many minutes in MVPA when this modeling was not applied.

An exercise diary was provided to each subject; they were asked to record up to four intentional bouts of exercise per day. This exercise diary included measures of start time, stop time, type of activity, and perceived exertion level.

IV. Bout Detection

A bout is defined as a period of time in which a person's PA level is at or above a specific intensity level for some duration of time. Please see Figure 1 for a detailed explanation.

A. Frequency and Duration of Bouts

A set of bouts was found for each subject and each data source, the ActiGraph Freedson (AG), Fitbit Intensity Score (FB), and modeled Fitbit Freedson (FF). To find each bout, we required that the PA level was elevated to a threshold of three (moderate PA) for a minimum length of ten minutes with a tolerance of a lower PA level of up to three minutes.

The number of bouts per day, as an average, was calculated for each subject and source. From this, the study mean and standard deviation was found. Likewise, a per subject per device mean duration was used to calculate the study mean and standard deviation of subject/source means.

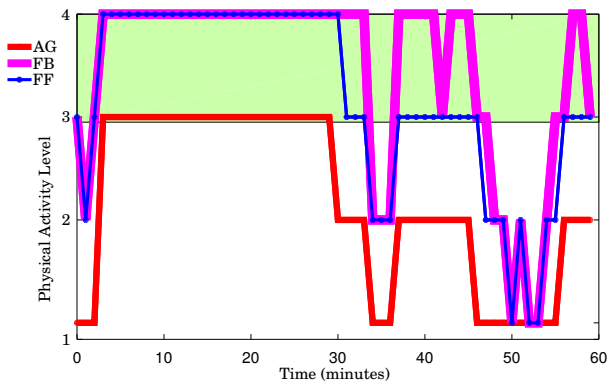

Fig. 1: As shown here, a bout can be defined as a period of time during which the PA level exceeds some threshold value. This figure shows two bouts, one starting at approximately the two minute mark and ending near the thirty minute mark, and the other starting near the thirty-five minute mark and ending ten minutes later. The start of a bout is visible after the fifty minute mark, but the end of that bout is not shown here.

A t-test was then performed, comparing the number of bouts per day from the AG to FB, and then again from AG to FF. A t-test was also performed to compare the bout durations, for all identified bouts, between the AG, FB, and FF data sources. A standard alpha of 0.05 was chosen to establish statistical significance.

B. Matching FB and FF Bouts to AG Bouts

The ActiGraph and Freedson VM3 method have been well established as a gold standard. As such, bouts found from the ActiGraph data source are here assumed to be the "truth."

We analyzed the PA bouts for each source to determine the accuracy of the FB and FF models compared to the ActiGraph GT3X model. Again, we began by evaluating each individual subject. The start time, stop time, and duration of each bout found from each source was saved to a spreadsheet. Matching was determined manually between the AG PA bouts and the FB and FF for each subject. A bout was considered a match if a majority of the minutes of one bout fell within another.

For each subject we found the difference between the stop and start time and the difference in the duration times of each model for every matched bout. Given this difference and the number of durations matched, we were able to calculate the average overestimation in duration time per bout of the FF and FB models compared to the AG results. We also summed the total duration minutes for each model and determined how many extra duration minutes were found in the FF model compared to the AG as well as how many extra duration minutes were found in the FB model compared to the AG. Finally we determined how many bouts were left unmatched in each model.

			vs AG	
	μ	σ	p	t
ActiGraph Freedson (AG)	0.34	0.47		
Fitbit Intensity (FB)	3.40	1.24	<.001	9.18
Fitbit Freedson (FF)	2.61	1.01	<.001	8.16

TABLE I: Number of bouts reported by each method, normalized across subjects per subject day. Also reported are the p and t values found from a t-test between the AG and FB results and AG and FF results.

We then compared the results of each model for all subjects combined. This included the average duration overestimation per bout of the FF model and FB model compared to the AG model along with the corresponding standard deviations. The total number of FB and FF extra duration minutes over the course of the entire study was summed and compared. The mean, standard deviation, and number of AG, FB, and FF bouts, for total and unmatched bouts, was found over all subjects and compared. Finally the average number of bouts per day and standard deviation for each model was calculated. With these calculations we were able to draw conclusions on both the FF and FB models compared to the AG model.

C. Self Report Diary

Our final analysis involved comparing the AG, FB, and FF bout results to those manually recorded by subjects in their diaries. We used the same method as described above to compare the identified bout start, stop, and duration times of each bout to that which the subjects had reported as intentional exercise. Self report bouts included the start and stop time, as well as the activity type logged. Knowing that subjects might report an inaccurate time, relative to when they may have actually started the exercise, the tolerance for matching was selected to be less stringent that the methods above. However, these were once again matched by hand to identify what best looked to be representative of each bout.

V. Results

A. Frequency and Duration of Bouts

Both the Fitbit Intensity and the Fitbit Freedson methods reported significantly ($p < 0.001$) more bouts than the ActiGraph Freedson method; this is shown in Table I. However, the duration of the bouts was found to be similar across each method ($p > 0.05$), see Table II.

B. Matching FB and FF to AG Bouts

The bouts identified from the Fitbit Intensity Score and from the Fitbit Freedson Modeling were matched to the most similar ActiGraph Freedson bouts as described above. Further, where two or more bouts were found in the FB or FF sources, of which all were within the time frame defined by AG, those bouts were combined to one representative bout instead. This comparison to the ActiGraph allows us to investigate the over- and under-estimation of the Fitbit based assessments as compared to the ActiGraph.

			vs AG*	
	μ	σ	p	t
ActiGraph Freedson (AG)	14.58	12.85		
ActiGraph Freedson (AG*)	21.20	9.65		
Fitbit Intensity (FB)	17.95	4.12	0.240	1.20
Fitbit Freedson (FF)	16.53	3.35	0.084	1.80

* adjusted to remove zero bout reports

TABLE II: The duration of bouts reported by each method, in minutes, averaged across the bouts detected per subject. Here the five subjects where AG reported zeros bouts for the duration of the data collection were removed in the adjusted (AG*) t-test.

	μ	σ	min	max
Different in duration per matched bout				
Fitbit Intensity (FB)	+15.55	15.13	-1	+48.33
Fitbit Freedson (FF)	+8.33	7.33	+0.33	+25.33
Sum of unmatched bout minutes per day				
Fitbit Intensity (FB)	+67.46	49.12	+16.8	+201.6
Fitbit Freedson (FF)	+33.86	15.09	+14.2	+63.1

TABLE III: Shown here are the mean differences in durations for matched bouts and the mean number of MVPA classified bouts per subject day as compared to the bouts identified in the ActiGraph Freedson data source.

Both the Fitbit Intensity and Fitbit Freedson methods were found to over estimate the duration of each matched bout in most all cases. As expected from the results in the previous section (V-A), both methods also resulted in an overestimate of minutes spent in MVPA bouts per day. These results are shown in Table III. While both measures over report, one should note that the modeling has an observable effect in reducing the per bout overestimate (15.55 minutes reduced to 8.33 minutes) and in reducing the per day unmatched bout minutes (67.46 reduced to 33.86 minutes).

C. Self Report Diary

As seen in Table IV, on average over 50% of the diary recordings are found in both the Fitbit Intensity (FB) and Fitbit Freedson (FF) reports, while an average of only 22.5% of the diary recordings are found in the ActiGraph Freedson (AG) reports. These averages correspond with the average number of bouts per subject also seen in Table IV. The results are not entirely surprising given that the average number of bouts per subject using the AG method is relatively small at 4 bouts/subject compared to the FB and FF methods at averages of 45.3 bouts/subject and 34.2 bouts/subject, respectively. We would expect the presence of a larger number of bouts to lead to a higher probability of finding a specific bout in a subset. It is important to take into account that these percentages may be slightly underrepresented due to the fact that some activities may not reflect as elevated MVPA with certain devices (e.g. stretching) and the possibility of user error when manually logging specific times for intentional PA.

We also find a large difference in the minimum and maximum average number of bouts per subject from the

	Number of Bouts				Fraction of Diary Found		
Subject	FF	AG	FB	SR	FF	AG	FB
1	18	0	18	5	100%	0%	100%
2	17	3	29	8	50%	25%	100%
3	13	8	11	8	37.5%	37.5%	62.5%
4	66	20	65	28	89.3%	64.3%	89.3%
5	21	2	31	5	40%	20%	40%
6	46	6	58	9	33.3%	33.3%	33.3%
7	18	0	28	5	0%	0%	60%
8	37	3	41	13	23.1%	23.1%	23.1%
9	33	9	56	15	53.3%	60%	60%
10	77	4	101	8	62.5%	37.5%	62.5%
11	13	0	19	7	28.6%	0%	28.6%
12	45	4	83	16	68.8%	25%	75%
13	38	2	42	16	62.5%	12.5%	62.5%
14	38	0	53	-	-	-	-
15	9	0	13	1	100%	0%	100%
16	58	3	76	5	60%	0%	80%
μ	34.2	4	45.3	9.9	53.9	22.5%	65.1%
σ	20.4	5.1	26.6	6.7	28.4	21.3%	25.9%
min	9	0	11	-	0	0%	23.1%
max	77	20	101	28	100	64.3%	100%

TABLE IV: The number of bouts identified per subject and per data source, regardless of the number of days recorded for each subject, is shown in the second through fifth columns. FF is the Modeled Fitbit Freedson, AG is the ActiGraph Freedson, FB is the Fitbit Intensity, and SR is the user provided Self Report. The last three columns show the percentages of the self report identified bouts of activity that were captured by each of the device/methods. The raw Fitbit Intensity score clearly captured the greatest number of self reported bouts (65% on average), but at the cost of also reporting a larger number of overall bouts (45). The ActiGraph on the other hand captured the least of the self reported bouts (23%), as also evidenced by the small average number of identified bouts (4). It can not be concluded from this data which of the AG, FB, or FF are most representative of the actual MVPA bouts, but it is visible that the modeled approach (FF) is able to provide an estimate between the two extremes.

AG to the FB and FF methods. It is interesting to note that the maximum number of bouts found using the AG method (20) is closer to the minimum number of bouts for both the FF (9) and FB methods (11) than each methods maximum number of bouts (77 and 101, respectively). We also find for multiple subjects that no bouts are detected using the ActiGraph method, which further establishes the significant differences in the calculated averages between the methods.

VI. Discussion and Conclusions

The ActiGraph GT3X and Fitbit devices are inherently very different devices. First, the ActiGraph GT3X has been designed and validated for applications where it is worn on the waist of the subject. While early Fitbit devices were also waist worn, the current product line focuses on wrist worn devices. This is likely a result of the adaptation and sales seen by Fitbit. Because the Fitbit devices easily synchronize their data to the Fitbit Cloud, do not require regular physical researcher access, are inexpensive, and have the ability to easily integrate into social media, they have high potential to be used for large scale physical activity assessment studies. However, they have historically lacked the congruence with observational studies and scientific efforts demonstrated by the many studies using ActiGraph GT3X devices. This paper sought to identify how modeling the PA level assessments of the ActiGraph using Fitbit data impacts bout assessment, and to inform future studies on the possibility of using such modeling on Fitbit data to utilize Fitbit devices at scale rather than the ActiGraph GT3X devices. One should note that this study has not considered all possible populations, and is not as such suggesting that the Fitbit devices are suitable replacements for the ActiGraph devices in all scenarios.

The overestimates shown in the number of bouts per subject per day, the duration of those bouts, and the sum of minutes in unmatched bouts is likely representative of the different if wear location. The wrist likely experiences more movement in a day than a person's waist, especially given how much time most Americans spend sitting. However, one should also be quick to note that the ActiGraph was unable to identify many self reported bouts that the Fitbit methods were both able to capture. This suggests that the overall sensitivity of the Fitbit, considering the hardware and the wear location, is higher than that of the ActiGraph. The discordance between self report and ActiGraph identified bouts also suggests that while the Freedson method is considered a gold standard for identification of minutes spent in MVPA, it may not be an ideal solution for estimating the times and durations of bouts of activities. A more detailed study investigating other proxy measures of activity such as heart rate may inform other methods of accurately identifying bouts of activity using a wearable device. Until then, it is suggested that researchers proceed with use of the Fitbit in clinical studies knowing that application of the modeling methods presented in other research by the authors is impactful in considering the time spent in MVPA, but not clinically equivalent in considering bouts of activity at MVPA levels.

References

[1] Susannah Fox and Maeve Duggan. Mobile Health 2012. *Pew Internet: Washington, D.C.*, page 29, 2012.

[2] Abby C. King, Karen Glanz, and Kevin Patrick. Technologies to Measure and Modify Physical Activity and Eating Environments. *American Journal of Preventive Medicine*, 48(5):630–638, 2015.

[3] Guides Fy and Non-gaap Gross Margin. Fitbit Reports $712M Q415 and $1.86B FY15 Revenue; Guides to $2.4 to $2.5B Revenue in FY16. *Fitbit*, 2016.

[4] Company News and Product News. Fitbit Leads the Wearables World : Canalys. pages 1–7, 2014.

[5] Gregory M Dominick, Kyle N Winfree, Ryan T Pohlig, and Mia A Papas. Physical Activity Assessment Between Consumer- and Research-Grade Accelerometers: A Comparative Study in Free-Living Conditions. *JMIR mHealth and uHealth*, 4(3):e110, 2016.

[6] Kyle N Winfree and Gregory Dominick. Modeling Clinically Validated Physical Activity Using Commodity Hardware (in review). *BHI 2017*.

[7] Jeffer E. Sasaki, Dinesh John, and Patty S. Freedson. Validation and comparison of ActiGraph activity monitors. *Journal of Science and Medicine in Sport*, 14(5):411–416, 2011.

[8] Patty S. Freedson, Edward Melanson, and John Sirard. Calibration of the Computer Science and Applications, Inc. accelerometer. *Medicine and science in sports and exercise*, 30(5):777–781, 1998.