Modeling Clinically Validated Physical Activity Using Commodity Hardware

Kyle N. Winfree, Gregory Dominick

Abstract— **Fitbit devices are one of the most popular wearable activity monitors in the consumer market. They are considerably cheaper than many of their clinical grade counterparts. However, they utilize proprietary algorithms for estimation of physical activity (PA). This study aims to model the measures of PA as reported by the ActiGraph GT3X using Fitbit measures of steps, METs, and intensity level. Such a model relating the Fitbit to what would have been reported by the ActiGraph could enable researchers to use the Fitbit instead of the ActiGraph in some applications, thus reducing cost or increasing the number of subjects involved in a study.**

This paper describes a study in which a model of the Freedson VM3 physical activity classification was constructed that uses measures from the Fitbit device instead of the typically provided ActiGraph vector magnitude. The data from 19 subjects, who concurrently wore both the ActiGraph and Fitbit devices for an average of 1.8 weeks, was used to generate the minute level based model. Several modeling methods were tested; a naïve Bayes classifier was chosen based on the lowest achieved error rate. That model reduces overall Fitbit to ActiGraph errors from 19.97% to 16.32%, a notable improvement. More importantly, it reduces the errors in moderate to vigorous physical activity levels by 40%, eliminating a statistically significant difference between MVPA estimates provided by the Freedson VM3 and Fitbit Intensity scores. This justifies use of the Fitbit device in place of an ActiGraph device in some large scale studies, especially those where MVPA estimates are of importance.

I. Introduction

Tri-axial accelerometers such as the ActiGraph GT3X are considered the criterion measure for objectively assessing physical activity (PA) within free-living environments [1], [2], [3]. However, research-grade accelerometers and the software required to analyze PA data are expensive and have limited applications beyond the scope of PA assessment.

Recent advancements in accelerometer technology have led to a proliferation of low-cost, wearable activity monitors, such as the Fitbit and Garmin Vivoactive devices; these are marketed directly to consumers and have seen highly successful adoption. Emerging evidence suggests that Fitbit may independently facilitate PA behavior change in the short-term [4], [5] and may be useful for promoting the current PA guidelines that recommend adults achieve 150 minutes of moderate-to-vigorous PA (MVPA) per week [6]. Many studies have found these inconsistencies in measurement congruence between Fitbit and research grade accelerometers, potentially limiting the utility of Fitbit within research settings and could have significant health implications if Fitbit users think 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. A successful model such as this would enable researchers to use the commercially available Fitbit devices in place of ActiGraph devices. Therefore, the primary aim of this study was to model validated measures of PA intensity as determined by ActiGraph GT3X accelerometers using Fitbit measures of intensity level, METs, and steps per minute.

II. Methods

The study design, procedures, instrumentation, and methods to assess PA have been previously described (7) and are briefly summarized here. Study participants consisted of a convenience sample of 19 healthy adult men (n = 4) and women (n = 15) who owned a Fitbit Flex device and were recruited from the University of Delaware. All volunteer participants provided their written informed consent and study approval was obtained by the University of Delaware Institutional Review Board. Participants provided their Fitbit username and password that were used to link individual Fitbit devices to a secure cloud-based server (Fitabase, Small Steps Labs), which enabled the continuous collection of participant steps, METs, and activity level, recorded in minute-intervals throughout the observed wear-periods. ActiGraph GT3X accelerometers were initialized using the ActiLife software (version 6.11.9) and set to record in 60-s epochs. Participants were instructed to wear the Fitbit Flex (wrist) and ActiGraph GT3X (waist) concurrently during all waking hours over 7 consecutive days. After completing the 7-day wear period, participants were asked to complete a second 7-day wear-period, for which most agreed (n = 16) and that occurred approximately three weeks after completing the first wear-period.

III. Instruments

The ActiGraph GT3X (ActiGraph, Pensacola, FL) is a small research-grade tri-axial accelerometer that is typically worn at the waist to provide objective measures of PA behavior in free-living conditions [7]. While several analysis methods are available to identify PA levels from raw measures, the standard in adult populations is the Freedson VM3 method used here [8]. This categorizes PA levels of sedentary, light, moderate, vigorous, and very vigorous activities from accelerometer counts detected within a specified time period. Intensity levels were defined based on the following cut-points: sedentary (< 200), light ($200 - 2690$), moderate ($2691 - 6166$), vigorous ($6167 - 9642$), and very vigorous

K. Winfree is an Assistant Professor in the Informatics and Computing Program at Northern Arizona University and is the contact author: kyle.winfree@nau.edu.

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

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

(> 9643). The vigorous and very vigorous categories were later combined to be consistent with activity data from the Fitbit Flex.

The Fitbit provides measures of steps per minute, METs, and a Fitbit proprietary intensity level. The intensity score follows the same concept as that of the Freedson VM3 algorithm; level 1 is sedentary, 2 is light activity, 3 is moderate, and 4 is vigorous.

IV. Modeling Intensity

In order to make it possible to make direct comparisons, we sought to develop a model of the Freedson VM3 results using measures provided by the Fitbit Flex.

A. Time registration of data from each device in Octave

ActiGraph and Fitbit data sets were combined in Octave using a set of custom programming functions. These functions compose an Octave (and Matlab) library, referred to as "WearWare[1]."

B. Freedson conversion from VM to PA Level

After the data registration was confirmed, the minute-level ActiGraph vector magnitude activity counts were converted to PA levels using the Freedson VM3 algorithm as implemented in the WearWare Toolkit library.

C. Modeling PA from Fitbit Measures

1) Making the classifier: We developed a machine learning classifier (i.e. model), which uses measures from the Fitbit, to predict intensity levels that the ActiGraph/Freedson VM3 would have estimated. To do this, we used the minute-by-minute ActiGraph/Freedson VM3 data as a gold standard [8], and trained four predictive models using the following classifier types: linear discriminant analysis classifier (LDA), quadratic discriminant analysis classifier (QDA), naïve Bayesian classifier (NBC), and a mahalanobis distance based classifier (MDA). A random subset of the entire data set, independent of the total number of minutes recorded from each subject, was used to identify the best predictive model. The distribution of the percentage of minutes in sedentary, light, moderate, and vigorous Freedson VM3 PA intensity used for the training sample was considered a very good representation of the entire data set (74.4%, 20.2%, 4.4%, 0.99% and 74.5%, 20.1%, 4.4%, 0.99%, respectively). The model with the best agreement rate (Cohen's kappa), as determined by testing on the training data only, was selected.

2) Selection Method Testing: To ensure that the most appropriate model was selected, the chosen model was retested using four different sampling methods. First, we utilized the 25% sampling approach from both weeks 1 and 2, but tested on the remaining 75% of data. Second, we used 100% of the week 1 data for training and tested on 100% of the week 2 data. Third, we selected a randomly equal maximal number of samples from each intensity level from week 1 (which resulted in 100% of the vigorous data being selected), and tested on 100% of the week 2 data. Fourth, as a check for over fit rather than any serious consideration of using the resulting model, we used 100% of week 1 and week 2 data for training and tested on the same 100% of week 1 and 2 data.

3) Identifying the Time Spent in Each Activity Level: Both the Fitbit and Freedson algorithms supply a minute by minute assessment of activity level, corresponding to sedentary (SED), light (LPA), moderate (MPA), and vigorous (VPA) levels. The developed classifier is intended to predict what the Freedson VM3 analysis on ActiGraph data would have provided, but solely from the Fitbit measures of steps per minute, METs, and the Fitbit intensity level. As such, each device, and the classifier analysis on the Fitbit device, provides an output that can be compared to the other. We performed an analysis to find the portion of time spent in each of the PA levels for the ActiGraph Freedson VM3 (labeled AG), Fitbit (FB), and Fitbit based Freedson Model (FF) at the group and the subject levels. To do this, we summed the number of cases where the PA level was equal to each condition of SED, LPA, MPA, and VPA and divided that by the number of observations (minutes) for that subject. To identify the representative group agreements and differences, we then found the mean and standard deviation across subjects for these percent of minutes in each PA level results. A sum of the mean MPA and VPA categories was used to find the group MVPA mean (from the group MPA and VPA means). A pooled standard deviation was used to find the SD for MVPA from the SD of MPA and VPA across subjects).

4) Identifying Agreement, Corrections, and Failures of the Predictive Model: To identify how the classifier is able to correct reflected PA levels in the Fitbit Intensity levels, we considered all combinations and cases where each device and the model reported each of the PA levels. There are a total of 64 combinations here, where AG = 1, FB = 1, FF = 1, ... AG = 1, FB = 4, FF = 2, etc. To track the reclassification of each minute, we calculated the percentage of minutes for each combination observed in the data set.

V. Results

A. Time Registration

Sixteen of the nineteen subjects were found to have an acceptable registration alignment without time shifting.

B. Freedson Conversion and Direct Comparison of Agreement to Fitbit Intensity

Prior to modeling, the direct ActiGraph/Freedson to Fitbit agreement of intensity levels was found to be 79.52% (Cohen's kappa). One should note that the distribution of PA levels in the data sets does have a strong influence on the agreement. The Freedson VM3 algorithm identified 74.50% of the total number of minutes across all subjects as sedentary, indicating that the simplest and most naïve approach of always labeling the PA as sedentary would be correct almost 75% of the time. However, this approach would lack any agreement for minutes in light, moderate, and vigorous intensity levels. Furthermore, we would be unable to determine activity bouts.

[1]https://www2.nau.edu/wearable-icp/index.php/projects/wearware/

C. Modeling PA from Fitbit Measures

1) Making the Classifier: Results were similar for the four classifier types (linear discriminant analysis, quadratic discriminant analysis, naïve Bayesian, and mahalanobis distance based classifier) used to test the training sample. Each produced an overall agreement of 82.82%, 83.24%, 83.75%, and 81.76% respectively. The naïve Bayes classifier (83.75% agreement) was chosen for further testing and development.

2) Selection Method Testing: The selection testing results demonstrated that selection method had minimal impact on the success of the classifier. Using the first method (training with 25% of the total data and testing on the remaining 75% of the sample), agreement (Cohen's kappa) for the training and testing samples were 83.70% and 83.77%, respectively. Results of the second testing method (training with all of the first week of data and testing on the second week only), were similar, with 84.20% agreement for the testing data and 83.68% agreement for the training data. One might expect that sampling a random and equal number of each PA level category, in attempt to remove any distribution bias, would improve the agreement on the testing set. However, agreement on the testing set was found to be 64.96%, while agreement when testing the training set (a much smaller set that the two previous methods) was still very similar, with an agreement rate of 82.11%. The final testing method involved selecting all data from all subjects for the training set, and testing again on the complete data set. This agreement from the model generated from this method was found to be 83.80%, approximately that of the first two methods.

3) Identifying the Time Spent in Each Activity Level: The primary goal of this paper was to develop a method for converting the Fitbit data into data more equivalent to that of the ActiGraph, thus enabling researchers to use the Fitbit in place of the ActiGraph in some applications. It is common in PA assessment to calculate the percentage of time spent in MVPA. Table I shows the placement of minutes in each PA level before the modeling is applied to the Fitbit data and the results of reclassification once the modeling is applied.

Table **??** shows the combined subject scores for all minutes observed. These are reported in percent of time. These results suggest that the Fitbit Intensity score reports almost twice the number of minutes in MVPA than the Freedson VM3 reports (9.99% vs 5.41%). The model corrected this by reducing the error overestimation to just 0.57%. This analysis was done at the subject level first and then combined into the group level.

4) Identifying Agreement, Corrections, and Failures of the Predictive Model: Considering all minutes used in this study and tracking the reclassification of those minutes, as shown in Figure 1, one can see the improvement in similarity between the ActiGraph Freedson and Fitbit Freedson estimates as compared to the ActiGraph Freedson and Fitbit Intensity similarities.

This analysis was done on the entire data set at the group level, not at the subject level. The greatest improvement was seen in the 2.29% of group minutes that Fitbit had classified as MPA, when the Freedson VM3 had classified them as LPA, which the model was able to bring back to the LPA category. Also notable are the cases where those minutes had been classified as MPA by the Freedson VM3, but Fitbit had classified them as VPA, and the Fitbit Freedson model reclassified them as MPA (1.86%). These are important, as they were counted towards minutes in moderate to vigorous activity (MVPA) levels by the Fitbit, but the model was able to correct these minutes to non-MVPA minutes.

AG minutes	AG(S) 251,152	AG(L) 67,890	AG(M) 14,743	AG(V) 3,344
FB(SED)	**94.51%**	44.13%	6.63%	4.84%
FB(LPA)	4.17%	**34.90%**	5.53%	2.72%
FB(MPA)	1.10%	19.93%	**35.20%**	11.96%
FB(VPA)	0.22%	1.04%	52.64%	**80.47%**
FF(SED)	**94.51%**	44.13%	6.63%	4.84%
FF(LPA)	4.96%	**49.38%**	17.11%	5.44%
FF(MPA)	0.47%	6.31%	**66.21%**	44.98%
FF(VPA)	0.07%	0.18%	10.06%	**44.74%**

TABLE I: Shown here is a confusion matrix identifying the percentage of the number of minutes in each paired category before and after classification. AG is the ActiGraph Freedson VM3, FB is the Fitbit Intensity, and FF is the modeled Fitbit Freedson. At first inspection, one might conclude that the Fitbit Freedson model performs poorly at vigorous levels. However, one should note that the Fitbit Freedson model reclassifies many of the minutes from moderate and vigorous PA levels into the light and moderate levels, increasing the agreement in those categories which also represent a greater number of minutes in the original data set. The Fitbit Freedson model is unable to reclassify minutes initially classified by the Fitbit Intensity score as sedentary (S).

PA	AG	FB FF	p	t
SED	74.41 ± 5.94	79.69 ± 5.16 79.69 ± 5.16	0.01 0.01	2.683 2.683
LPA	20.17 ± 4.75	10.32 ± 2.83 14.33 ± 4.18	$*$ $*$	7.131 3.697
MPA	4.37 ± 1.76	6.47 ± 2.71 4.98 ± 2.32	0.01 0.41	2.591 0.837
VPA	1.04 ± 1.12	3.53 ± 1.81 1.00 ± 1.61	$*$ 0.94	4.669 0.081
MVPA	5.41 ± 1.48	9.99 ± 2.31 5.98 ± 1.99	$*$ 0.37	6.690 0.919

$*$ Statistically significant difference, $p < 0.01$.

TABLE II: This table shows the percentage of time spent in each PA level as classified by each of the devices/algorithms. AG is the ActiGraph Freedson VM3, FB is the Fitbit Intensity, and FF is the modeled Fitbit Freedson. Note that the percentage of time in MVPA is reported by the Fitbit Intensity score to be nearly twice that which was determined by the ActiGraph with the Freedson VM3 method. However, the Fitbit Freedson model is able to largely correct this error, making the reported difference no longer significantly different.

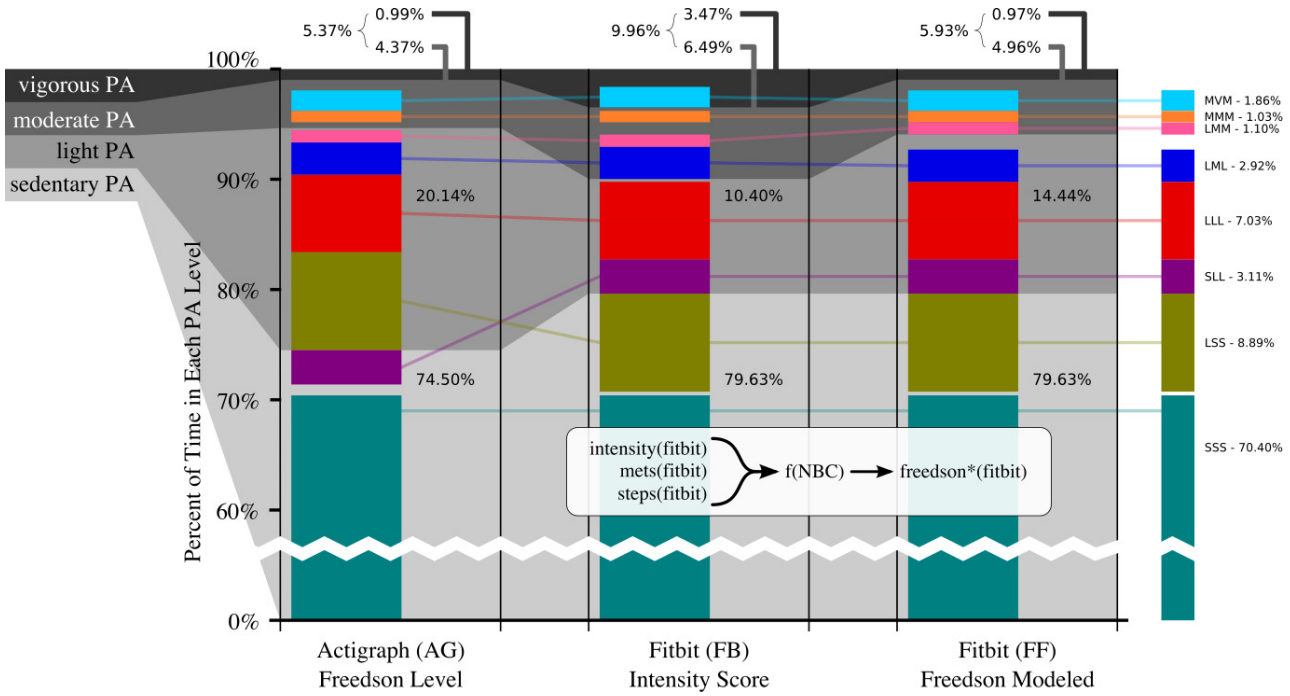

Fig. 1: Tracking the change in per-minute classification of PA level. Shown are the per device/algorithm estimated number of minutes in each PA level (sedentary, light, moderate, and vigorous) in various shades of gray. Overlaid above this, in color, are the number of minutes meeting each category level across all three methods. For example, 70.40% of the minutes measured were assessed as sedentary by all three methods (SSS), while 1.86% of the minutes measured were assessed as moderate, vigorous, and moderate PA by the ActiGraph/Freedson, Fitbit Intensity, and Fitbit Freedson methods respectively. Minute categories that were smaller than 1% of the data are not shown in the interest of space.

VI. Discussion and Conclusions

The ActiGraph and Fitbit each report activity level using very different processing algorithms. Through use of the Fitbit reported measures of steps, METs, and intensity level, one is able to predict with reasonable success the ActiGraph Freedson VM3 reported PA levels in 83.68% of the data collected in this study $(\Sigma(SSS, LLL, LML, MVM, < 1\%\ agreements))$. Considering assessed PA levels of agreement between the Freedson VM3 and Fitbit only, the devices had a 80.03% agreement rate (this includes several categories not reflected in Figure 1 due to the small size, $< 1\%\ agreements$). Most importantly, this model is able to correct the total time spent in moderate to vigorous PA levels (MVPA) to a measure no longer statistically significantly different from that which was reported by the ActiGraph Freedson VM3 model, a method considered to be the gold standard for PA assessment in the ecologically valid community setting.

One possible reason for the remaining differences between the ActiGraph and Fitbit methods is the placement of the device on the wearer. The ActiGraph was worn on the subjects' waists, while the Fitbit was worn on their wrists. A wrist worn device though, is expected to have greater adherence in large subject pools due to the convenience and unobtrusiveness of the devices.

A further study will also investigate the observed bouts of activity as measured by each of these devices and methods.

References

[1] Nancy F. Butte, Ulf Ekelund, and Klaas R. Westerterp. Assessing physical activity using wearable monitors: Measures of physical activity. *Medicine and Science in Sports and Exercise*, 44(SUPPL. 1):5–12, 2012.

[2] John Dinesh and Patty Freedson. Actigraph and Actical Physical Activity Monitors:A Peek Under the Hood. *Med Sci Sports Exerc*, 44:1–6, 2012.

[3] Richard P. Troiano, David Berrigan, Kevin W. Dodd, Louise C. Mâsse, Timothy Tilert, and Margaret Mcdowell. Physical activity in the United States measured by accelerometer. *Medicine and Science in Sports and Exercise*, 40(1):181–188, 2008.

[4] Lisa A. Cadmus-Bertram, Bess H. Marcus, Ruth E. Patterson, Barbara A. Parker, and Brittany L. Morey. Randomized Trial of a Fitbit-Based Physical Activity Intervention for Women. *American Journal of Preventive Medicine*, 49(3):414–418, 2015.

[5] Julie B Wang, Lisa A Cadmus-Bertram, Loki Natarajan, Martha M White, Hala Madanat, Jeanne F Nichols, Guadalupe X Ayala, and John P Pierce. Wearable Sensor/Device (Fitbit One) and SMS Text-Messaging Prompts to Increase Physical Activity in Overweight and Obese Adults: A Randomized Controlled Trial. *Telemedicine and e-Health*, 21(10):1–11, 2015.

[6] William L. Haskell, I. Min Lee, Russell R. Pate, Kenneth E. Powell, Steven N. Blair, Barry A. Franklin, Caroline A. MacEra, Gregory W. Heath, Paul D. Thompson, and Adrian Bauman. Physical activity and public health: Updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. *Medicine and Science in Sports and Exercise*, 39(8):1423–1434, 2007.

[7] 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.

[8] 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.