Modeling Clinically Validated Physical Activity Assessments Using Commodity Hardware

Kyle N. Winfree and Gregory Dominick

***Abstract*—Consumer-grade wearable activity devices such as Fitbits are increasingly being used in research settings to promote physical activity (PA) due to their low-cost and widespread popularity. However, Fitbit-derived measures of activity intensity are consistently reported to be less accurate than intensity estimates obtained from research-grade accelerometers (i.e., ActiGraph). As such, the potential for using a Fitbit to measure PA intensity within research contexts remains limited. This study aims to model ActiGraph-based intensity estimates from the validated Freedson vector magnitude (VM3) algorithm using measures of steps, metabolic equivalents, and intensity levels obtained from Fitbit. Minute-level data collected from 19 subjects, who concurrently wore the ActiGraph GT3X and Fitbit Flex devices for an average of 1.8 weeks, were used to generate the model. After testing several modeling methods, a naïve Bayes classifier was chosen based on the lowest achieved error rate. Overall, the model reduced Fitbit to ActiGraph errors from 19.97% to 16.32%. Moreover, the model reduced misclassification of Fitbit-based estimates of moderate-to-vigorous physical activity (MVPA) by 40%, eliminating a statistically significant difference between MVPA estimates derived from ActiGraph and Fitbit. Study findings support the general utility of the model for measuring MVPA with the Fitbit Flex in place of the more costly ActiGraph GT3X accelerometer for young healthy adults.**

***Index Terms*—Activity recognition, error correction, wearable sensors.**

I. Introduction

RESEARCH grade, tri-axial accelerometers such as the ActiGraph GT3X, quantify acceleration forces measured in one or more orthogonal axes (x, y, z or vertical, antero-posterior, and medio-lateral depending on wear location and orientation) and are considered the criterion measure for objectively assessing physical activity (PA) within free-living environments [1]–[3]. The composite sum of activity counts derived from all three orthogonal axes is commonly used to calculate the vector magnitude (VM3) for a user-defined time sampling interval (i.e., epoch). ActiGraph data are typically processed using proprietary software (i.e., ActiLife) and validated algorithms to define intensity thresholds (i.e., cut-points) for specific populations (e.g., toddlers, adults, older adults) and estimate time spent in sedentary, light, and moderate-to-vigorous physical activity (MVPA) [4]. 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 Fitbits which are marketed directly to consumers and have demonstrated widespread acceptability and adoption. Coinciding with national estimates that indicate approximately 60% of adults track their weight, diet, or exercise, and approximately 45% of American adults report owning at least one activity tracker, up from 21% in 2014 [5], [6]. These consumer-based devices have helped promote the “quantified-self” movement [7]. With 29 million registered users worldwide, Fitbit is the most popular brand of wearable activity tracking monitor sold on the commercial market [8], [9]. In addition to having an integrated tri-axial accelerometer, Fitbit devices wirelessly sync processed data to a users mobile Fitbit application (i.e., Fitbit app) and to a personal Fitbit web account (i.e., Fitbit dashboard), enabling users to continuously monitor and track various attributes of their activity and activity bouts in real-time. Furthermore, Fitbit incorporates several known behavior change techniques that are designed to reinforce behavioral self-monitoring such as goal-setting and behavioral feedback provision [10], all within an interactive, user-friendly application.

Emerging evidence suggests that Fitbit may independently facilitate PA behavior change in the short-term [11], [12] and may be a useful tool for promoting the current national PA guidelines, which recommend that adults achieve 150 minutes of MVPA per week [13]. However, to date, PA research has relied on ActiGraph accelerometers to quantify MVPA outcomes due to the limited evidence regarding the accuracy of Fitbit-derived intensity estimates. For example, previous research has demonstrated that Fitbit devices are more likely to significantly overestimate MVPA compared to ActiGraph accelerometers in free living conditions [13]–[15]. Moreover, the error estimates for Fitbit-derived intensity classification are shown to increase for specific bouts of activity such as walking, running, and weight lifting [13], [16]. 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

Manuscript received May 15, 2017; revised September 22, 2017 and November 15, 2017; accepted December 19, 2017. Date of publication December 25, 2017; date of current version March 5, 2018. This work was supported by a grant from the Arizona Biomedical Research Center, ADHS16-00005488, Arizona New Investigator Award. *(Corresponding author: Kyle N. Winfree.)*

K. N. Winfree is with the School of Informatics, Computing, and Cyber Systems, Northern Arizona University, Flagstaff, AZ 86011 USA (e-mail: kyle.winfree@nau.edu).

G. Dominick is with the University of Delaware, Newark, DE 19716 USA (e-mail: gdominic@udel.edu).

Digital Object Identifier 10.1109/JBHI.2017.2787461

significantly overestimated the percent of time spent in sedentary, moderate, and vigorous PA by 26%, 3%, and 3% per day, respectively [15]. In contrast, one study reported that MVPA estimates were not significantly different between the Fitbit Flex and ActiGraph GT3X+. However, study findings were based on only one day of data [17]. Other attempts to compare intensity estimates between Fitbit and ActiGraph have also been reported. Alharbi and colleagues [18] used metabolic equivalent tasks (METs) to estimate minutes of MVPA in a sample of cardiac rehabilitation patients over a 4 day period. Whereas the Fitbit Flex was shown to overestimate MVPA by 10 min/day, the authors also reported that Fitbit was able to identify participants who achieved 150 minutes of MVPA per week (positive predictive value = 0.98, 95% CI = 88.23; 99.94). Despite this finding, reliance on MET values as determined by Fitbit is shown to be significantly different compared to ActiGraph derived METs for overall daily activity and minutes of self-reported exercise [13]. However, these findings may be biased as the authors relied on MET values that were determined by Fitbit, which others have found to be significantly different compared to ActiGraph GT3X for overall daily activity and minutes of self-reported exercise [15].

Based on the current literature, it appears that the proprietary algorithms Fitbit uses to define intensity cut-points are not consistent with the validated algorithms researchers use with ActiGraph accelerometers, resulting in disparate MVPA estimates. This "cut-point non-equivalence" impedes measurement comparisons between devices. In contrast to the validated ActiGraph algorithms that are available to researchers via the ActiLife software, the proprietary algorithms used by Fitbit have yet to be clinically validated to date. Given that increasing weekly minutes of MVPA remains a key outcome for PA interventions and epidemiological studies, it is vital that researchers rely on valid objective measures of PA behavior. Moreover, as Fitbit is consistently reported to overestimate minutes of MVPA [15], [18], [19], this could have broader effects in terms of public health, as Fitbit users may believe they are meeting the national PA guidelines when in fact, they are not.

As a commercial device, Fitbit has proprietary ownership of the technology and algorithms used to quantify activity and to date, researchers are unable to access the raw acceleration data collected on the Fitbit device. This closed-source approach further inhibits measurement transparency and limits measurement comparisons. 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 recent study by Hickey and colleagues used this approach to create cut-points for the wrist-worn ProDiary activity monitor that were based on raw ActiGraph GT3X accelerometer counts sampled at 30 Hz within 1-minute epochs [20]. Yet, no existing studies have modeled PA levels as reported from an ActiGraph accelerometer using only Fitbit data. If successful, a model such as this would enable researchers to use the commercially available Fitbit devices in place of ActiGraph devices, which could significantly reduce study costs.

This paper builds on a previous paper presented by Drs. Winfree and Dominick [21]. Two devices were used in this study. The first was the clinically validated ActiGraph GT3X [22]. Data from these devices can be scored using a variety of validated algorithms in order to classify sedentary behavior and categories of activity intensity (light, moderate, vigorous, and MVPA) [23]. The second device was the Fitbit Flex. While some Fitbit devices have been validated in small, targeted studies, these have largely focused on step counts involving walking or running bouts assessed in controlled clinical settings [24]–[26].

This study aimed to model the Fitbit to the Freedson PA levels using at least one week of continuous data in which subjects concurrently wore the Fitbit Flex and ActiGraph GT3X device during all waking hours within free-living conditions with the higher level objective of developing a model to relate PA intensity as determined by ActiGraph GT3X accelerometers to measure of intensity, METs, and steps per minute from the Fitbit device.

II. Methods

The study design, procedures, instrumentation, and methods to assess PA have been previously described [15] and are briefly summarized here. Study participants consisted of a convenience sample of 19 healthy adult men (n = 4) and women (n = 15) between the ages of 19 and 37 years old, 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.

During the baseline measurement visit, participants completed a standard demographics survey and had their height (cm) and weight (kg) recorded via stadiometer and digital scale, respectively. These measures were used in the subjects' device profiles. 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),[1] 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

A. ActiGraph GT3X

The ActiGraph GT3X (ActiGraph, Pensacola, FL)[2] 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 [15]. Although the Freedson VM3 algorithm is commonly used in the literature to classify activity intensity thresholds such as time spent in light, moderate, and vigorous intensity PA [23], it has only been validated for healthy adult

[1] http://www.fitabase.com/

[2] http://actigraphcorp.com/

populations and may be less accurate for defining sedentary behavior. Despite these limitations, researchers continue to rely on ActiGraph MVPA estimates as the "gold-standard" of objective PA measurement. This algorithm is well established for use with healthy adults, and serves as an appropriate choice given that the study sample consisted of young, healthy adults.

Activity data from the ActiGraph GT3X accelerometers were processed within the ActiLife software. Daily wear-times were validated using the Troiano (2007) algorithm [3] in which the device was worn for at least 8 hours/day. Non-wear periods were defined if no epoch counts were recorded for 60 or more continuous minutes. Intensity levels were defined based on the following cut-points: sedentary (<200), light ($200 - 2690$), moderate ($2691 - 6166$), vigorous ($6167 - 9642$), and very vigorous (>9643). The vigorous and very vigorous categories were later combined to be consistent with activity data from the Fitbit Flex.

As described earlier [15], minute-level episodes of inactivity were defined in the following way: VM count <200 + step count $= 0$, based on previously established criteria [27].

B. Fitbit Flex

The Fitbit device displays continuous measures of daily step counts, distance traveled, energy expenditure, and active minutes (defined as $\geq$10 minutes spent in continuous activity, considered to be $\geq$ to 3 METs (i.e., moderate intensity PA).[3] When viewing the Fitbit Dashboard, a user can monitor daily totals and averages over time. However, from a measurement perspective, these day-level totals or averages limit the types of analyses that can be performed. The ability to consider minute-by-minute level data is expected to enable researchers to ask different questions over the coming few years than they have been able to the past few years.

Fitbit stores minute-level user data on the Fitbit Cloud which is available via the Fitbit Application Programming Interface (API), or through a fee-for-service third party such as Fitabase,[4] which also utilizes the Fitbit API. Fitabase enables collection of these minute level measures that are otherwise unavailable from the Fitbit Dashboard alone. Alternatives to Fitabase, such as that developed by Kitsiou should consider other aspects concerning data management such as complying with federal HIPAA regulations for ensuring patient privacy [28].

Classification of Fitbit intensity levels is likely similar to the cut-point approach that the Freedson VM3 algorithm uses. However, intensity levels are categorized ordinally; level 0 is sedentary (or non-wear), 1 reflects light activity, 2 and 3 represent moderate and vigorous intensity activity, respectively. These categories were re-coded to 1, 2, 3, and 4, all respectively, to match the coded values reported by the Freedson VM3 method.

ActiGraph and Fitbit data were time registered to ensure that minute-level measures from both devices were consistent. Fitbit non-wear periods were determined if a 0 was recorded for >60 continuous minutes. Data were excluded if either ActiGraph or Fitbit indicated periods of non-wear. Self-reported exercise bouts were considered valid if the days and minutes of reported exercise were within five minutes of the activity counts that were concurrently measured with ActiGraph.

Shown in Fig. 1 is a high level pictorial representation of the data flow and processes of data collection, model generation, and model application used in this study.

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."[5] While WearWare now has a similar data collection feature to that which Fitabase offers, that aspect was not used during this study. A script was used to identify individual ActiGraph AGD files while simultaneously searching for all corresponding Fitbit files, matched by date and time (minute-level) into a struct format.

A registration check was performed for each subject during the data loading process to ensure that the synchronization was successful. Data registration was done by selecting the intersection of the two sets where each specific minute datum was present in both sets.

The registration was shifted by five minutes in each direction, one minute at a time, and the cumulative sum of the differences in steps per minute was again calculated. A ratio of the difference in cumulative steps to the total number of steps recorded by Fitbit was used to establish a "goodness of fit" threshold. When two sets are correctly registered at the start, this ratio is found to be low at the zero shift point. Likewise, the ratio should increase for each minute of shift, in each direction, forming a "V" shape around the zero shift point.

B. Freedson Conversion From VM3 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. This has been implemented in the Wear Ware Matlab/Octave library and follows the method outlined on the ActiGraph website and [23].

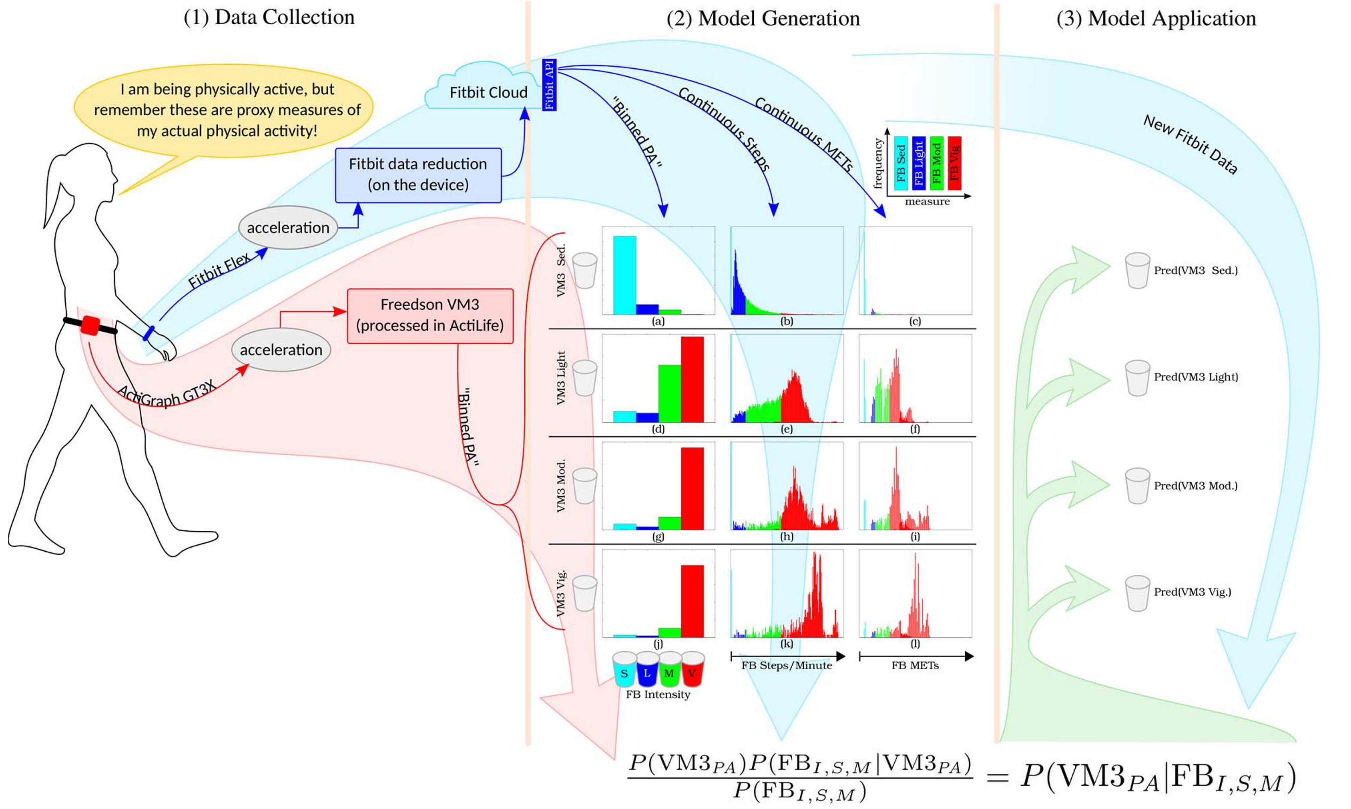
C. Modeling PA From Fitbit Measures

1) Making the Classifier: We developed a statistical 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

[3] Fitbit website: What are active minutes: https://help.fitbit.com/articles/en_US/Help_article/1379

[4] www.fitabase.com

[5] https://www.nau.edu/winfreelab



$$\frac{P(\mathrm{VM3}_{PA})P(\mathrm{FB}_{I,S,M}|\mathrm{VM3}_{PA})}{P(\mathrm{FB}_{I,S,M})} = P(\mathrm{VM3}_{PA}|\mathrm{FB}_{I,S,M})$$

Fig. 1. Shown here are the (1) data collection, (2) model generation, and (3) model application processes of this work. (1) Data collection happens concurrently, with measures from the wrist (Fitbit Flex) and waist (ActiGraph GT3X), each in an effort to accurately capture the activity of the wearer. Both the ActiGraph GT3X and Fitbit Flex provide proxy measures and estimates of a user/wearer's actual activity. As such, the divergent placement of each device (wrist vs. hip), is less significant so long as both make provide similar unit-measures of activity. (2) The model generation step then considers both similarities and difference between the measures from each device. Here, the ActiGraph GT3X and Freedson VM3 algorithm are considered the criterion measure. As the distribution of Fitbit to ActiGraph agreement demonstrate, measurement congruence varies between devices across intensity classifications. Yet, measurement congruence can be improved by considering Fitbit measures of steps/minute or estimated METs. This is apparent when considering the relative distributions of each Fitbit measure (column) for the given ActiGraph measure (row). For example, consider the case in the top left plot. This plot shows the distribution of Fitbit Intensity Levels for the case where the VM3 method categorized this data as sedentary (Sed). As one might expect, and is shown in the top middle plot, the distribution of steps/minute is biased towards a low number of steps/minute than other VM3 categories. Then considering the Fitbit Intensity Level distributions for ActiGraph-VM3 estimates of MVPA, one can see that the Fitbit steps/minute and METs are biased toward different ends of the scale, exemplifying the differences that can be utilized to inform prediction of the VM3 category. (3) The final step is model application. Here, new measures from the Fitbit are used as inputs, along with the statistical findings from the model generation step, to predict what physical activity levels would have been assessed by the VM3 method had an ActiGraph GT3X also been present.

gold standard [23], 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 shifting the data sources. A well aligned and a poorly aligned sample are shown in Fig. 2.

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.

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.

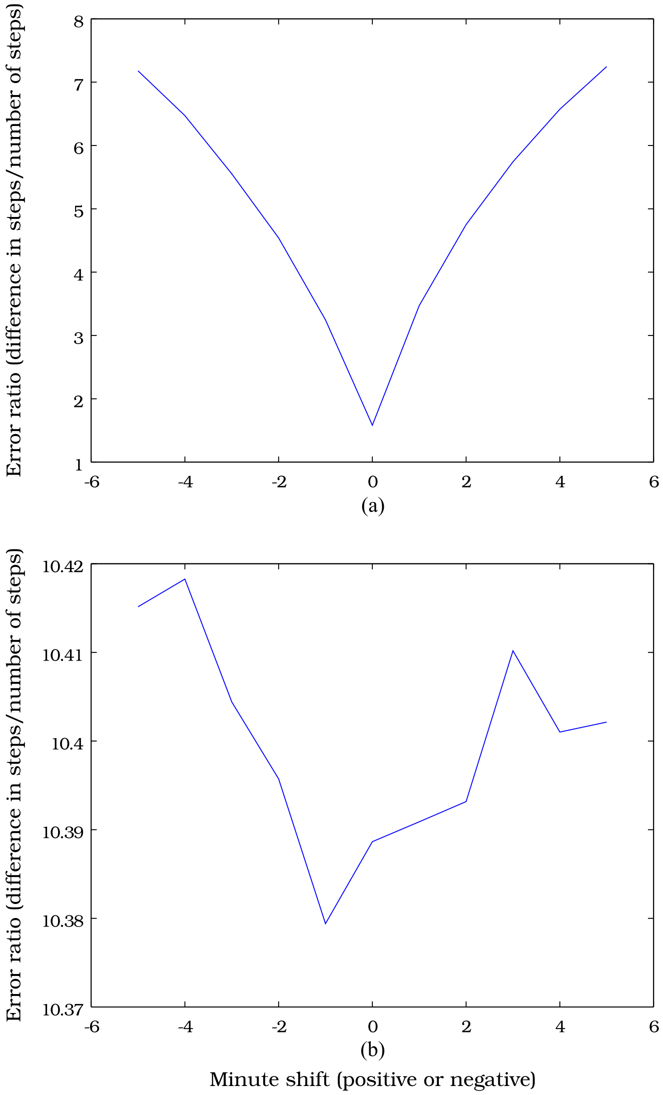


Fig. 2. (a) An ideal time synchronization. This is apparent from the sharp "V" shape, at which the minimum is located at the point of zero time shift and the small ratio (≈ 1.5) around this point. (b) A poor time synchronization. This can been seen from the large ratio (> 10) of the difference in cumulative steps to the total number of steps recorded.

An included Appendix provides an explanation of the mathematics used to apply the statistical learning classifier developed here. One may also consider using the Octave integrated function `test_sc.m`.

2) Naïve Bayes and Available Information: The naïve Bayes classifier considers each feature (Fitbit PA level, steps/minute, and METs) to contribute independently to the probability of class label (VM3 PA level), regardless of the correlations between features (Fitbit PA level, steps/minute, and METs). As such, this classifier does not assume the case of no correlation between features. Instead, each feature is unique enough to provide insights that help inform predictive class label. To demonstrate how reducing the available information to the classifier impacts the results, we generated naïve Bayes classifier using subsets of all available information: only steps/minute, METs, or Fitbit Intensity, and the paired cases between these. A

TABLE I
THIS TABLE ILLUSTRATES HOW THE USE OF AVAILABLE INFORMATION IMPACTS THE QUALITY OF THE NAÈVE BAYES CLASSIFIER

	AG(S) = FF(S)	AG(L) = FF(L)	AG(M) = FF(M)	AG(V) = FF(V)	ε	Comment
I, steps/minute, METs	94.50%	85.12%	2.32%	59.86%	0.0670	Baseline for other comparisons.
I, steps/minute	94.51%	87.72%	-	5.54%	0.0672	Failure at (M).
I, METs	94.50%	87.64%	0.62%	32.84%	0.0665	Better at (L), at cost to (V).
I	94.50%	87.86%	-	-	0.0674	Failure at (M) and (V).
steps/minute, METs	96.79%	81.84%	3.16%	43.11%	0.0464	Better (S), at cost to (L), (M), and (V).
steps/minute	97.47%	80.10%	1.64%	-	0.0423	Failure at (V).
METs	98.26%	75.35%	4.92%	13.38%	0.0361	Best (S), at cost to (L), (M), and (V).

$*$ I = Fitbit Intensity, (S) = sedentary, (L) = light PA, (M) = moderate PA, (V) = vigorous PA. Each row reflects the diagonal axis components of a larger confusion matrix. Values reported are the percent of VM3 classified data captured correctly for each corresponding modeled PA level. For example, the columns in the first row show the percent of physical activity classified by VM3 as sedentary that was also classified with the Fitbit Freedson (FF) as sedentary, percent of VM3 light and FF light, percent of VM3 moderate and FF moderate, and finally percent of VM3 vigorous and FF vigorous for the case where all available information was used to inform the classifier. The next row reflects the same measures for the case where only steps/minute were used, and so on. Where there is a - present in the table represent cases where the model was unable to classify any data into this category.

table illustrating only the diagonal axis agreement rates of each of these cases is provided in Table I. The researchers felt that the model which made complete use of available data provided the best balance of success across all classifications. As shown though, the overall error rate (ε) cannot be used as the end all measure of model success. Here, the lowest error rate is reflected for the condition where only METs are used to classify the PA level, resulting in a high success rate for sedentary classification at the cost of the other categories. Such a detriment to the other categories makes this such classifier unusable in research. The model that includes Fitbit Intensity, steps/minute, and METs may not have the lowest error rate, but it does however capture the best representation of the four PA classes.

3) 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.

4) 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 II shows the placement of minutes in each PA level before the modeling

TABLE II
SHOWN HERE IS A CONFUSION MATRIX IDENTIFYING THE PERCENTAGE OF THE NUMBER OF MINUTES IN EACH PAIRED CATEGORY BEFORE AND AFTER CLASSIFICATION

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

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).

TABLE III
THIS TABLE SHOWS THE PERCENTAGE OF TIME SPENT IN EACH PA LEVEL AS CLASSIFIED BY EACH OF THE DEVICES/ALGORITHMS

PA	AG	FB FF	p	t
1	74.41 ± 5.94	79.69 ± 5.16	0.01	2.683
		79.69 ± 5.16	0.01	2.683
2	20.17 ± 4.75	10.32 ± 2.83	$*$	7.131
		14.33 ± 4.18	$*$	3.697
3	4.37 ± 1.76	6.47 ± 2.71	0.01	2.591
		4.98 ± 2.32	0.41	0.837
4	1.04 ± 1.12	3.53 ± 1.81	$*$	4.669
		1.00 ± 1.61	0.94	0.081
MVPA	5.41 ± 1.48	9.99 ± 2.31	$*$	6.690
		5.98 ± 1.99	0.37	0.919

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

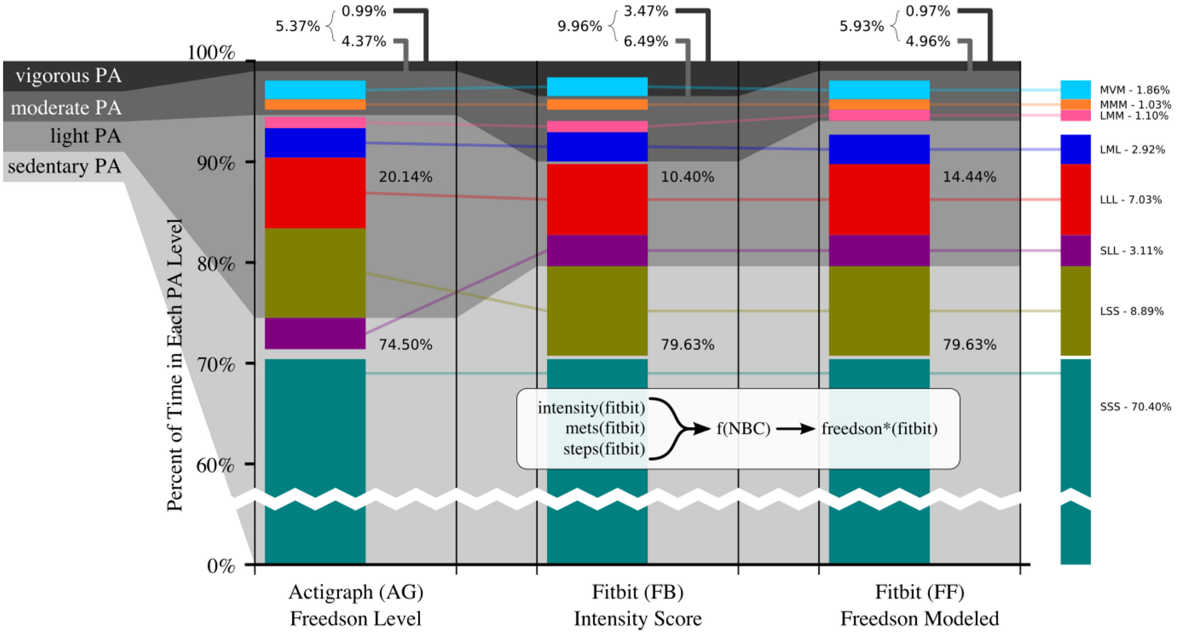


Fig. 3. Tracking the change in per-minute classification of PA level. Shown are the per device/algorithm estimated number of minutes in each PA level (sedentary, light, moderate, and vigorous) in various shades of gray. Overlaid above this, in color, are the number of minutes meeting each category level 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.

is applied to the Fitbit data and the results of reclassification once the modeling is applied.

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.

5) Identifying Agreement, Corrections, Failures, and Independence of the Predictive Model: Considering all minutes used in this study and tracking the reclassification of those minutes, as shown in Table III, 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.

Independence of the model, from such factors as height, weight, or BMI of the subject, was evaluated by correlation of these factors with the model agreement to the Freedson model PA level (as % fit). This is shown in Fig. 4, where as it is very difficult to discern a pattern that relates these factors. The largest correlation found was with weight of the subject, -0.22. However, this correlation is still much lower than the typically accepted "small" correlation of 0.7 for known dependence. Height and BMI show an even lower correlation to model fit at -0.6 and -0.16 respectively. The authors expect that this is due to the fact that both devices, the Fitbit Flex and the ActiGraph GT3X require an input of subject height, weight, sex, and age before use. Though the researchers cannot say for sure, we expect that both are using these measures to calibrate measurement to the descriptors of the wearer. As such, we expect that both devices have accounted for these measures before providing output, thus eliminating the expectation for use to see subject descriptor influence in the generated model.

The model was shown to be dependent on the percent time in moderate to vigorous physical activity level, as determined by the uncorrected Fitbit intensity level. This correlation is shown in Fig. 5. Also shown in this figure is the correlation and dependence of congruence between Fitbit intensity and the ActiGraph based PA level, as a function of the MVPA time. Prior to application of the model, the Fitbit intensity was found to have a congruence of approximately 70%–90%, with a lower congruence at higher percent time MVPA. When considering the model though, this congruence was improved to approximately 90%–97%, with again a lower congruence at higher percent time in MVPA. One can use these linear correlations to estimate the expected errors in both original Fitbit assessed PA levels and modeled PA levels, both based solely on the Fitbit assessed time in MVPA. Most notably, the average congruence between the

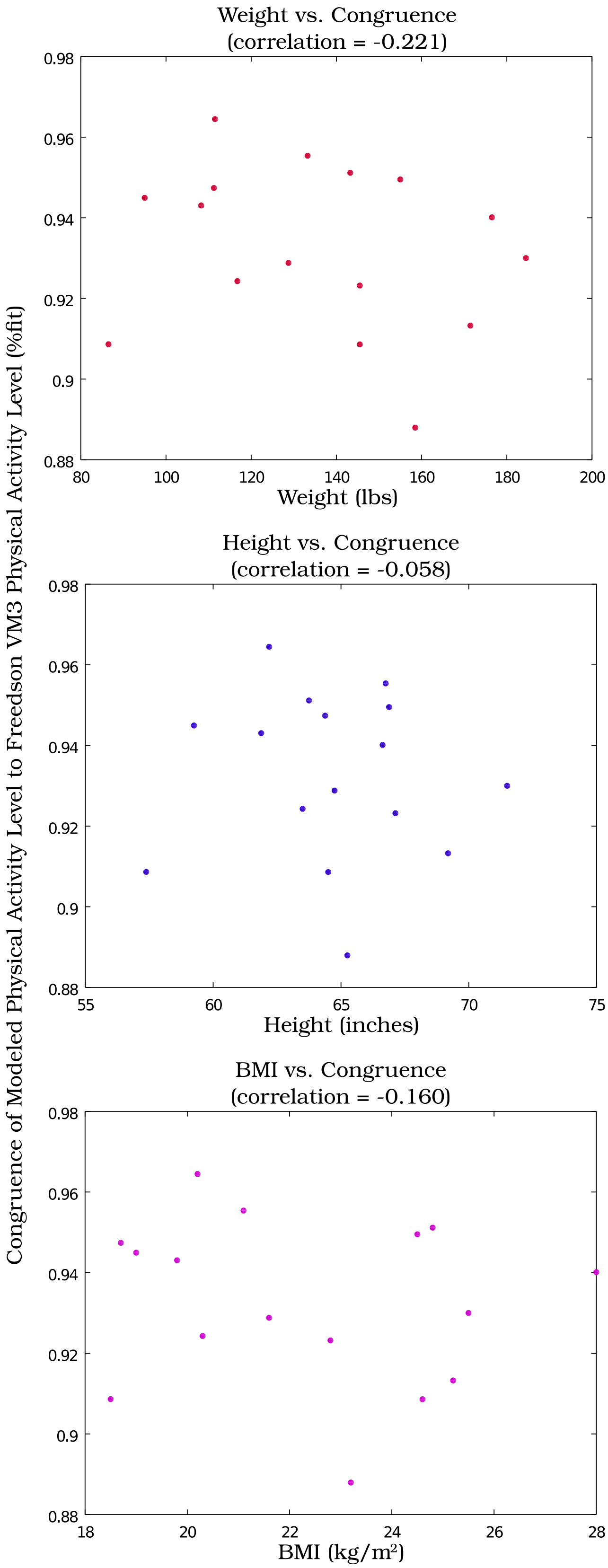


Fig. 4. This set of figures demonstrates how weight, height, and BMI have no observable effect in the overall accuracy of the model. Each of these factors presents a very low correlation to the model congruence, suggesting that the model congruence is independent of the subject's weight, height, or BMI.

Fitbit intensity and ActiGraph Freedson PA level was 80%, a rate unacceptable for clinical use of the Fitbit in place of the ActiGraph. However, by application of the modeling presented in this paper, that average congruence is increased to 93%, a point which is arguably within an acceptable range for use of the Fitbit in some studies.

VI. Discussion

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 Fig. 3 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. An over estimation of time spent in MVPA could be expected to be detrimental in determining the success of interventions. Given the data at hand, one is also able to estimate the errors in the model relative to the PA level assessed by the ActiGraph. This, shown in Fig. 5, also demonstrates that the model is most accurate for subjects who are less active, suggesting that the model would be well suited for use in applications where one is studying MVPA of largely sedentary subjects.

Not only have we outlined how to improve the MVPA estimates from the Fitbit Flex, we have also demonstrated the ability to collect data remotely. This contrasts with the traditional protocol where participants must return the data collection to the researcher for the researcher to manually download data from the GT3X device. Future longitudinal studies can leverage these tools that enable single in-person interactions with the participant, while continuously collecting measures of physical activity; participant contact can become minimized and can extend for the life of the device. The authors have developed the WearWare server system, utilizing the Fitbit API, and are in the pilot process of collecting data without third party engagements.

A. Limitations

One possible reason for the remaining differences between the ActiGraph and Fitbit is the placement of the device on the wearer; the Freedson VM3 algorithm was validated while ActiGraph device was worn at the waist. The Fitbit was worn on the wrist, and may be more susceptible to measurement error. However, the algorithms used in conjunction with ActiGraph have only been validated when the device is worn at the hip. Despite this measurement limitation, results of this study demonstrate the extent to which Fitbit-derived data can be improved, particularly MVPA, which is most often the primary outcome measure

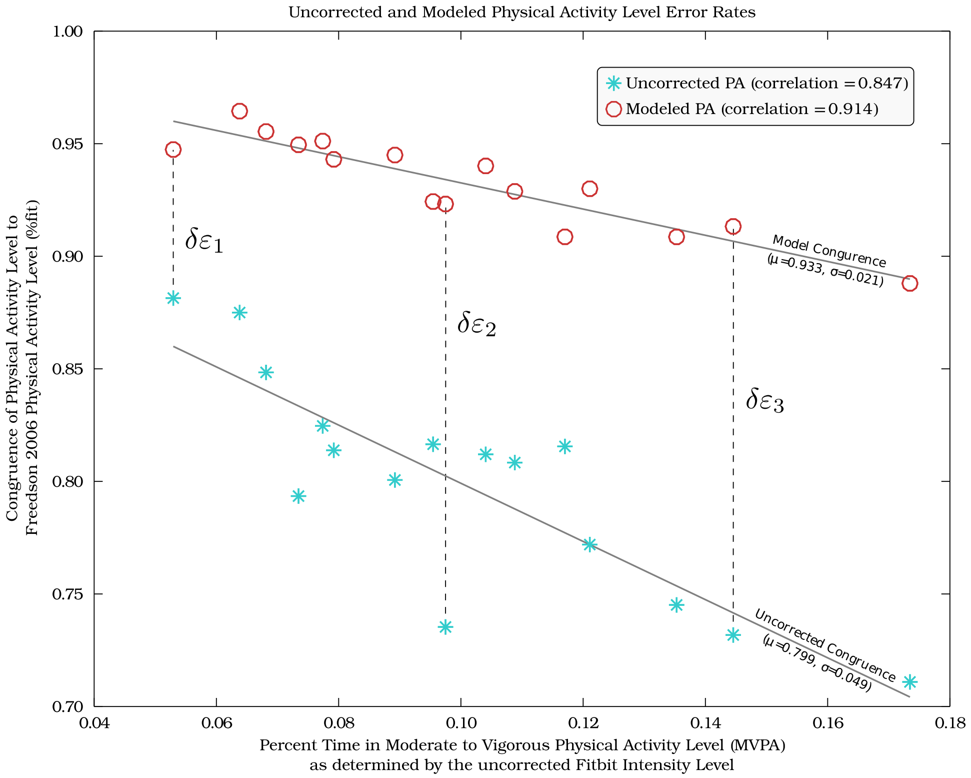


Fig. 5. Prior to modeling, the Fitbit was shown to have an error rate that substantially depended on the time spent in MVPA. At higher MVPA, these errors could be expected to be as great as 30%; 20% on average. However, with the application of the modeling, such error rates are reduced to 10% at most, or approximately 7% on average. One should also note the change in slope between the unmodeled and modeled lines here. While the modeled PA does still depend on the subject's time in MVPA, this dependence is not as strong as that established by the unmodeled PA. This is evidenced by the difference in slope, where a horizontal line would be reflective of no dependence of congruence on activity level. The degree of error reduction does vary between subjects, though as illustrated in Fig. 4, not predictably by height, weight, or BMI. Three examples are shown here, $\delta\varepsilon_1$, $\delta\varepsilon_2$, and $\delta\varepsilon_3$. For each of these three subjects, the Fitbit estimated time in MVPA was dramatically different. The similarity of error reduction between $\delta\varepsilon_2$ and $\delta\varepsilon_3$, and difference between these and $\delta\varepsilon_1$ demonstrates the variable reduction in error, specific to each participant.

used in interventions and epidemiological surveillance studies that assess PA.

Wearable activity monitors, including research-grade accelerometers are constantly evolving and most have now transitioned to being worn on the wrist given that wrist-worn devices are shown to result in greater wear-time adherence due to the convenience and unobtrusiveness of the devices.

The work presented here makes the assumption that the ActiGraph GT3X, in conjunction with the Freedson VM3 method, is able to accurately reflect the actual physical activity of the involved participants. Instead of relying on the measures assessed by the ActiGraph, future research should consider direct observation of the maximal oxygen consumption to establish participant physical activity level. A future study will also work to consider if other measures from the Fitbit, such as the recent advent of photoplethysmography (PPG) to monitor heart rate, can be leveraged to improve the estimates of PA using these low-cost commodity devices.

B. Conclusions

We now clarify the contribution of this work. Given the importance of measuring MVPA in physical activity research as an outcome variable and considering the fact that the majority of literature supports the understanding that Fitbit overestimates PA, reliance on Fitbit reported PA estimates could be a detriment to determining the success of an intervention. The degree of error one should expect were outlined in Fig. 5. These errors approach 30% with physically active subjects. Errors are magnified as actual MVPA increases. These errors can potentially have major implications in interpreting MVPA outcomes if using a Fitbit without applying data corrections. For example, one could consider the case where MVPA increases a small amount as result of an intervention, but yet the reported MVPA increase is large, and thus an intervention might appear to have statistically significant impact when the reflected change is not actually large enough to trigger significance. The provided statistical model for correction of Fitbit PA assessment makes use of three measures provided by Fitbit, steps, METs, and Intensity Level, considering the relative probabilities of each Freedson based PA level assessment given the Fitbit measures, rather than creating a cut point like model from Fitbit data. Fitbit does not currently provide raw measures (the accelerations) from their devices, making such a cut point based algorithm unfeasible.

Finally, the authors would like to note that they have prepared the necessary functions to apply these techniques to existing Fitbit data as a library that can be used within Octave or Matlab. The details of this library, including how to request gratis access, is available on Dr. Winfree's laboratory website. Likewise, the de-identified data used to construct this model is also available by request.

Appendix

The naïve Bayes classifier is a fairly simplistic probability classification model, which is based on the application of Bayes' theorem and independence between features. One can think of the simplest Bayesian model as

$$posterior = \frac{prior \times likelihood}{evidence},$$

or

$$p(C_k|\vec{x}) \quad = \frac{p(C_k)p(\vec{x}|C_k)}{p(\vec{x})}, \tag{1}$$

where C is the set of k class assignments, C_k; $\vec{x}$ is the feature, or measures used to predict class assignment; the $prior$, $p(C_k)$, is the probability of a specific class assignment without regards for the features specific to each class; the $likelihood$, $p(\vec{x}|C_k)$, is the probability of the evidence $\vec{x}$ given the class assignment C_k; the $evidence$, $p(\vec{x})$, is the probability of the given feature observation; and finally $posterior$, $p(C_k|\vec{x})$, is the conditional probability of C_k given the evidence $\vec{x}$ at hand.

In the case outlined in this paper, the class assignment set is defined as the physical activity level as assigned by the ActiGraph method [23],

$$\begin{aligned} C &= [sedentary, light\,PA, moderate\,PA, vigorous\,PA] \\ &= [1, 2, 3, 4]. \end{aligned} \tag{2}$$

The vector $\vec{x}$ is a feature comprised of measures from the Fitbit device,

$$\vec{x} = [steps(FB), METs(FB), intensity(FB)]. \tag{3}$$

However, the above simplification must be considered in the context of a multinomial event model. In this case, $p(\vec{x}|C_k)$ becomes

$$p(\vec{x}|C_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod p_{ki}^{x_i}. \tag{4}$$

To estimate the physical activity level that would have been assigned by the ActiGraph/Freedson method, but using the Fitbit measures alone, we will use a set of measures established by the training of the classifier. These include:

$$[\boldsymbol{CC_{mean}}] = \begin{bmatrix} \mu_{1,1} & \mu_{1,2} & \mu_{1,3} \\ \mu_{2,1} & \mu_{2,2} & \mu_{2,3} \\ \mu_{3,1} & \mu_{3,2} & \mu_{3,3} \\ \mu_{4,1} & \mu_{4,2} & \mu_{4,3} \end{bmatrix} \tag{5}$$

$$[\boldsymbol{CC_{var}}] = \begin{bmatrix} \sigma_{1,1}^2 & \sigma_{1,2}^2 & \sigma_{1,3}^2 \\ \sigma_{2,1}^2 & \sigma_{2,2}^2 & \sigma_{2,3}^2 \\ \sigma_{3,1}^2 & \sigma_{3,2}^2 & \sigma_{3,3}^2 \\ \sigma_{4,1}^2 & \sigma_{4,2}^2 & \sigma_{4,3}^2 \end{bmatrix} \tag{6}$$

$$\vec{CC}_n = \begin{bmatrix} n_{1,sedentary} \\ n_{2,lightPA} \\ n_{3,moderatePA} \\ n_{4,vigorousPA} \end{bmatrix} \tag{7}$$

where the number of columns in $[\boldsymbol{CC_{mean}}]$ and $[\boldsymbol{CC_{var}}]$ is equal to, and corresponds to the number of Fitbit measures used. The number of rows in these three matrices is equal to and corresponds to the number of physical activity levels, or the number of classes possible with the classifier. The specific values for each of these matrices, along with the supporting Octave library, are available upon request.

With these model parameters in hand, we can move to use of the multinomial naïve Bayes classifier to predict the class labels, physical activity level, using a new set of Fitbit measures $[\boldsymbol{X}]$. In this matrix, each row of $[\boldsymbol{X}]$ is a separate observation minute with columns represented measures of $steps$, $METs$, and Fitbit reported $intensity$. Linearizing Equations (1) and (4), we can use the following approach.

For each class k,

$$\begin{aligned} [\boldsymbol{z}] = ([\boldsymbol{X}] - [\mu_{k,1}\ \mu_{k,2}\ \mu_{k,3}])^2 \oslash [\sigma_{k,1}^2\ \sigma_{k,2}^2\ \sigma_{k,3}^2] + ... \\ ln([\sigma_{k,1}^2\ \sigma_{k,2}^2\ \sigma_{k,3}^2]), \end{aligned} \tag{8}$$

and

$$\vec{d}_{:,k} = \sum_{columns} \left(\frac{-[\boldsymbol{z}]}{2} \right) + ln\,(n_k)\,, \tag{9}$$

then

$$[\boldsymbol{D}] = e^{\left([\boldsymbol{d}] - ln(\vec{CC}_n)\right)} - \frac{ln(2 \cdot \pi)}{2}. \tag{10}$$

From which we then find the maximum value in each row. The column of this maximum is representative of the assigned class, as it exhibits the highest probability given the $prior$, $likelihood$, and $evidence$. These functions are also integrated into the Octave `test_sc.m` function for "classifier:nbc." To use this function, one needs to provide a struct with $[\boldsymbol{CC_{mean}}]$, $[\boldsymbol{CC_{var}}]$, and $\vec{CC}_n$, $[\boldsymbol{X}]$, no mode parameter, and a flag to use the "classifier:nbc" method internally. The $test\_sc()$ function is able to support many other classifier types, none of which were found to be as effective in this application.

References

[1] N. F. Butte, U. Ekelund, and K. R. Westerterp, "Assessing physical activity using wearable monitors: Measures of physical activity," *Med. Sci. Sports Exercise*, vol. 44, no. Suppl. 1, pp. 5–12, 2012.

[2] J. Dinesh and P. Freedson, "Actigraph and Actical physical activity monitors: A peek under the hood," *Med. Sci. Sports Exercise*, vol. 44, pp. S86–S89, 2012.

[3] R. P. Troiano, D. Berrigan, K. W. Dodd, L. C. Mâsse, T. Tilert, and M. Mcdowell, "Physical activity in the United States measured by accelerometer," *Med. Sci. Sports Exercise*, vol. 40, no. 1, pp. 181–188, 2008.

[4] P. S. Freedson, E. Melanson, and J. Sirard, "Calibration of the Computer Science and Applications, Inc. accelerometer," *Med. Sci. Sports Exercise*, vol. 30, no. 5, pp. 777–781, 1998.

[5] PwC, "The Wearable Life 2.0: Connected living in a wearable world," Los Angeles, CA, USA, pp. 1–23, May 2016.

[6] S. Fox and M. Duggan, "Mobile health 2012," Pew Internet, Washington, DC, USA, 2012, p. 29.

[7] A. C. King, K. Glanz, and K. Patrick, “Technologies to measure and modify physical activity and eating environments,” *Amer. J. Preventive Med.*, vol. 48, no. 5, pp. 630–638, 2015.

[8] 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 Inc., San Francisco, CA, USA, 2016.

[9] Company News and Product News, “Fitbit leads the wearables world: Canalys,” 2014, pp. 1–7.

[10] E. J. Lyons, Z. H. Lewis, B. G. Mayrsohn, and J. L. Rowland, “Behavior change techniques implemented in electronic lifestyle activity monitors: A systematic content analysis,” *J. Med. Internet Res.*, vol. 16, no. 8, 2014, Art. no. e192.

[11] L. A. Cadmus-Bertram, B. H. Marcus, R. E. Patterson, B. A. Parker, and B. L. Morey, “Randomized trial of a Fitbit-based physical activity intervention for women,” *Amer. J. Preventive Med.*, vol. 49, no. 3, pp. 414–418, 2015.

[12] J. B. Wang *et al.*, “Wearable sensor/device (Fitbit one) and sms text-messaging prompts to increase physical activity in overweight and obese adults: A randomized controlled trial,” *Telemed. e-Health*, vol. 21, no. 10, pp. 782–792, 2015.

[13] W. L. Haskell *et al.*, “Physical activity and public health: Updated recommendation for adults from the American College of Sports Medicine and the American Heart Association,” *Med. Sci. Sports Exercise*, vol. 39, no. 8, pp. 1423–1434, 2007.

[14] R. E. R. Reid *et al.*, “Validity and reliability of Fitbit activity monitors compared to ActiGraph GT3X+ with female adults in a free-living environment,” *J. Sci. Med. Sport*, vol. 20, no. 6, pp. 578–582, 2017.

[15] G. M. Dominick, K. N. Winfree, R. T. Pohlig, and M. A Papas, “Physical activity assessment between consumer- and research-grade accelerometers: A comparative study in free-living conditions,” *JMIR mHealth uHealth*, vol. 4, no. 3, 2016, Art. no. e110.

[16] Y. Bai *et al.*, “Comparison of consumer and research monitors under semistructured settings,” *Med. Sci. Sports Exercise*, vol. 48, no. 1, pp. 151–158, 2016.

[17] A. Sushames, J. G. Z. van Uffelen, and K. Gebel, “Do physical activity interventions in Indigenous people in Australia and New Zealand improve activity levels and health outcomes? A systematic review,” *Int. J. Behav. Nutrition Phys. Activity*, vol. 13, no. 1, 2016, Art. no. 129.

[18] M. Alharbi, a. Bauman, L. Neubeck, and R. Gallagher, “Validation of Fitbit-Flex as a measure of free-living physical activity in a community-based phase III cardiac rehabilitation population,” *Eur. J. Preventive Cardiology*, vol. 23, pp. 1476–1485, 2016.

[19] 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. Nutrition Phys. Activity*, vol. 12, 2015, Art. no. 42.

[20] A Hickey *et al.*, “Estimating cut points: A simple method for new wearables,” *Maturitas*, vol. 83, pp. 78–82, Jan. 2016.

[21] K. N. Winfree, G. Dominick, M. Papas, and R. Pohlig, “Modeling clinically validated physical activity using commodity hardware,” in*Proc. IEEE Int. Biomed. Health Informat. Conf.*, 2015, pp. 157–160.

[22] J. H. Migueles *et al.*, “Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations,” *Sports Med.*, vol. 47, pp. 1821–1845, 2017.

[23] J. E. Sasaki, D. John, and P. S. Freedson, “Validation and comparison of ActiGraph activity monitors,” *J. Sci. Med. Sport*, vol. 14, no. 5, pp. 411–416, 2011.

[24] 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. Sci. Med. Sport*, vol. 17, no. 5, pp. 496–500, 2013.

[25] K. M. Diaz *et al.*, “Fitbit®: An accurate and reliable device for wireless physical activity tracking,” *Int. J. Cardiology*, vol. 185, pp. 138–140, 2015.

[26] A. K. Singh, C. Farmer, M. L. E. Van Den Berg, M. Killington, and C. J. Barr, “Accuracy of the FitBit at walking speeds and cadences relevant to clinical rehabilitation populations,” *Disability Health J.*, vol. 9, pp. 320–323, 2015.

[27] N. Aguilar-Farías, W. J. Brown, and G. M. Peeters, “ActiGraph GT3X+ cut-points for identifying sedentary behaviour in older adults in free-living environments,” *J. Sci. Med. Sport*, vol. 17, no. 3, pp. 293–299, 2014.

[28] S. Kitsiou *et al.*, “Development of an innovative mHealth platform for remote physical activity monitoring and health coaching of cardiac rehabilitation patients,” in *Proc. 2017 IEEE EMBS Int. Conf. Biomed. Health Informat.*, 2017, pp. 133–136.