J Appl Physiol 124: 1284–1293, 2018.
First published January 25, 2018; doi:10.1152/japplphysiol.00760.2017.

RESEARCH ARTICLE

Cross-validation and out-of-sample testing of physical activity intensity predictions with a wrist-worn accelerometer

Alexander H. K. Montoye,[1,2] Bradford S. Westgate,[3] Morgan R. Fonley,[3] and Karin A. Pfeiffer[4]

[1]*Department of Integrative Physiology and Health Science, Alma College, Alma, Michigan;* [2]*Clinical Exercise Physiology Program, Ball State University, Muncie, Indiana;* [3]*Department of Mathematics and Computer Science, Alma College, Alma, Michigan; and* [4]*Department of Kinesiology, Michigan State University, East Lansing, Michigan*

Submitted 24 August 2017; accepted in final form 19 January 2018

Montoye AH, Westgate BS, Fonley MR, Pfeiffer KA. Cross-validation and out-of-sample testing of physical activity intensity predictions with a wrist-worn accelerometer. *J Appl Physiol* 124: 1284–1293, 2018. First published January 25, 2018; doi:10.1152/japplphysiol.00760.2017.—Wrist-worn accelerometers are gaining popularity for measurement of physical activity. However, few methods for predicting physical activity intensity from wrist-worn accelerometer data have been tested on data not used to create the methods (out-of-sample data). This study utilized two previously collected data sets [Ball State University (BSU) and Michigan State University (MSU)] in which participants wore a GENEActiv accelerometer on the left wrist while performing sedentary, lifestyle, ambulatory, and exercise activities in simulated free-living settings. Activity intensity was determined via direct observation. Four machine learning models (plus 2 combination methods) and six feature sets were used to predict activity intensity (30-s intervals) with the accelerometer data. Leave-one-out cross-validation and out-of-sample testing were performed to evaluate accuracy in activity intensity prediction, and classification accuracies were used to determine differences among feature sets and machine learning models. In out-of-sample testing, the random forest model (77.3–78.5%) had higher accuracy than other machine learning models (70.9–76.4%) and accuracy similar to combination methods (77.0–77.9%). Feature sets utilizing frequency-domain features had improved accuracy over other feature sets in leave-one-out cross-validation (92.6–92.8% vs. 87.8–91.9% in MSU data set; 79.3–80.2% vs. 76.7–78.4% in BSU data set) but similar or worse accuracy in out-of-sample testing (74.0–77.4% vs. 74.1–79.1% in MSU data set; 76.1–77.0% vs. 75.5–77.3% in BSU data set). All machine learning models outperformed the euclidean norm minus one/GGIR method in out-of-sample testing (69.5–78.5% vs. 53.6–70.6%). From these results, we recommend out-of-sample testing to confirm generalizability of machine learning models. Additionally, random forest models and feature sets with only time-domain features provided the best accuracy for activity intensity prediction from a wrist-worn accelerometer.

NEW & NOTEWORTHY This study includes in-sample and out-of-sample cross-validation of an alternate method for deriving meaningful physical activity outcomes from accelerometer data collected with a wrist-worn accelerometer. This method uses machine learning to directly predict activity intensity. By so doing, this study provides a classification model that may avoid high errors present with energy expenditure prediction while still allowing researchers to assess adherence to physical activity guidelines.

Address for reprint requests and other correspondence: A. H. K. Montoye, 614 W. Superior, Alma, MI 48801 (e-mail: montoyeah@alma.edu).

artificial neural network; decision tree; GENEActiv; random forest; support vector machine

INTRODUCTION

Accelerometers are an objective and increasingly popular tool used to measure physical activity (PA) (23). Traditional accelerometer use typically employs hip-worn devices that collect manufacturer-specific, proprietary "activity counts"; these counts are then used to estimate activity intensity via counts/minute thresholds, called "cut-points" (44). While initial evidence showed that activity counts and cut-points could be used with high accuracy for predicting energy expenditure and activity intensity for ambulation in a laboratory setting (11), follow-up studies found that cut-point methods were specific to the population, activities, setting, and accelerometer brand tested (29, 30, 39, 41, 43).

With recent technological improvements, the use of high-frequency raw data sampling and advanced analytic techniques (e.g., machine learning) has expanded the potential for accelerometers to be used for PA assessment. One appealing accelerometer placement is the wrist; studies using wrist-worn accelerometers have seen substantially improved compliance compared with hip-worn accelerometers (42), presumably because the wrist is a more comfortable and/or more familiar site to wear a device. Additionally, wrist-worn accelerometers collect movement data of the upper body, which is not possible with hip-worn accelerometers and which may offer additional information to characterize energy expenditure or activity intensity or even recognize specific types of PA being performed (10, 24, 27, 37).

Unfortunately, studies developing machine learning techniques often suffer from limitations similar to studies developing cut-points, i.e., models are developed in controlled laboratory settings and are not tested on independent samples. Studies performing such out-of-sample tests typically show reduced accuracy for PA assessment, similar to cut-point approaches (5, 12, 18). Recently, several studies have focused on improving generalizability of machine learning models. Kerr, Ellis, and colleagues used a wrist-worn accelerometer and a machine learning model to predict PA in five activity types (sitting, driving, standing, walking/running), and the machine learning model performed with >80% accuracy in both validation and out-of-sample testing in a diverse, adult sample (10, 15). Another study by Sasaki et al. used a wrist-

worn accelerometer and developed a machine learning model to classify five activity types (sedentary, standing, household, locomotion, recreational) but achieved only 49–61% accuracy in a free-living setting (36).

Such PA type prediction models have two notable limitations. First, there seems to be an inverse relationship between a model's accuracy and the number of PA types it can predict, so a practitioner must choose between accuracy and granularity for distinguishing activities. Second, PA type models must characterize all activities into one of the classes that they were developed to recognize. For example, the model developed by Kerr and Ellis classifies all activities as sitting, driving, standing, or walking/running (10, 15). Free-living individuals perform far more varied activities than these five, so it is possible that large portions of individuals' days could not be correctly classified with such a model.

An alternate approach is to classify activities directly according to their intensity. All activities, regardless of activity type, can be characterized into one of four activity intensities: sedentary, light, moderate, or vigorous. Additionally, activity intensity classification is congruent with most health-focused PA recommendations, which advocate accumulating a certain amount of time in moderate- or vigorous-intensity activities while reducing time spent sedentary (31). Activity intensity classification can be accomplished indirectly by first predicting the energy expenditure of a task and then classifying intensity based on energy expenditure thresholds, but this approach has been problematic because of both relatively high error and bias toward the mean (24, 37). With classification models, it is possible to bypass energy expenditure and directly estimate activity intensity. A preliminary study by members of our research group found >90% accuracy in classifying activity intensity as sedentary, light, or moderate-to-vigorous from an accelerometer worn on the left wrist (28). While this study provides evidence that direct classification of activity intensity may improve measurement accuracy over first predicting energy expenditure, this study was limited in that it only used one type of machine learning model, did not differentiate moderate- from vigorous-intensity activity, and was not tested in an independent data set. Therefore, the present study's purpose was to develop multiple types of machine learning methods for predicting activity intensity from a wrist-worn accelerometer and compare their performance in cross-validation within a single data set vs. a true out-of-sample test on an independent data set.

MATERIALS AND METHODS

Participants. This study utilized two previously collected data sets, one from Michigan State University (MSU) and one from Ball State University (BSU). The original studies were approved by their respective organization's institutional review board, and all participants provided written informed consent before enrollment. The principal investigators for both projects are authors of this report and received approval from their respective institutions to merge de-identified data from the two studies to use for analysis.

The MSU data set consisted of 39 apparently healthy adults aged 18–35 yr who were able to complete activities of vigorous intensity, including cycling and jogging. The BSU data set consisted of 24 apparently healthy adults aged 18–79 yr without orthopedic limitations who could safely perform self-paced jogging or fast walking as well as less intense activities. Demographics of the samples can be seen in Table 1. Participants in the BSU data set were, on average, older and of higher body mass index than participants in the MSU data set.

Table 1. *Demographic data from Michigan State University and Ball State University data sets*

	All	Male	Female
Michigan State University data			
n	39	19	20
Age, yr	22.1 (4.3)	23.7 (5.0)	20.6 (2.8)
Height, cm	171.4 (10.1)	179.1 (7.7)	164.1 (5.7)
Weight, kg	72.4 (16.2)	84.5 (13.1)	60.8 (8.9)
Body mass index, kg/m^2	24.4 (3.6)	26.3 (3.4)	22.5 (2.6)
Ball State University data			
n	24	12	12
Age, yr	46.3 (19.2)	48.7 (19.5)	43.9 (19.4)
Height, cm	174.0 (8.7)	179.9 (7.1)	168.1 (5.8)
Weight, kg	79.6 (15.5)	88.6 (13.0)	70.5 (12.4)
Body mass index, kg/m^2	26.1 (3.6)	27.3 (3.2)	24.9 (3.6)

Data are means (SD).

Equipment. In both data sets, time-stamped activity types were recorded by research assistants using Observerware (Educational Consulting, Hobe Sound, FL) software on a handheld tablet computer. Additionally, research staff fitted participants with GENEActiv accelerometers (ActivInsights, Kimbolton, UK) on the left wrist, which was the nondominant wrist for the majority of both samples. For the MSU data set, the GENEActiv monitors were initialized to record raw, triaxial data at a sampling rate of 20 samples/s (20 Hz). For the BSU data set, the GENEActiv monitors were initialized to record raw, triaxial data at 60 Hz.

Protocols. The protocols for both data sets took place in a simulated free-living setting within a research laboratory. For the MSU data set, participants performed 13 activities for 3–10 min each in a 90-min protocol. For the BSU data set, participants performed 12–21 activities for 2–15 min each in an 80-min protocol. Additionally, participants in the BSU data set had to choose at least four activities from each of three categories (sedentary, household, and ambulatory/exercise) and were asked to spend at least 50% of the protocol engaged in sedentary activities, to replicate an adult's typical day (7, 21). For both data sets, all activities were self-paced and the activity order, duration, and method to perform the activity were chosen by participants. Protocols were designed in this manner to replicate the freedom and conditions available to individuals in free-living settings while ensuring efficient data collection and utilizing high-quality criterion measures available in a laboratory setting. The activities, their categorization, and their intensity (determined a priori) are shown in Table 2. During the protocols, trained research assistants directly observed and recorded the exact start and stop times of all activities on a handheld tablet computer. Activities were each given a unique code, which was entered by research staff within 1 s of the start of the activity. Because of the free-living nature of the protocols and the short durations of the activities performed, direct observation and associated metabolic threshold (MET) predictions from the Compendium of Physical Activities, along with standard absolute MET thresholds (≤1.5: sedentary, 1.6–2.9: light, 3.0–5.9: moderate, and ≥6.0: vigorous), were used as the criterion measure for activity intensity.

Data processing and model development. Data collected from the two data sets were processed in the same way. Acceleration signals from the GENEActiv monitors among each of the three axes were divided into nonoverlapping 30-s intervals. Intervals with only one activity coded by the research assistant were used for model development and analysis, but intervals with two or more activities were coded as transitions and removed from the data set. Also, intervals spent moving from one activity to the next were coded as transitions and removed, as they could not be assigned an intensity.

Table 2. *Activity types, categories, and intensities performed in the protocols*

Activity Type	Activity Category	Activity Intensity	Compendium Code
	Michigan State University data		
Lying down	Sedentary	Sedentary	07010
Reading quietly	Sedentary	Sedentary	07070
Using a computer	Sedentary	Sedentary	07021
Standing	Household	Light	07041
Sweeping	Household	Light	05011
Folding laundry	Household	Light	05090
Walking slowly (overground)	Ambulatory/exercise	Light	17152
Walking briskly (overground)	Ambulatory/exercise	Moderate	17190
Jogging (overground)	Ambulatory/exercise	Vigorous	12020
Cycling (stationary, 50–100 W)	Ambulatory/exercise	Moderate	02017
Stair use	Ambulatory/exercise	Vigorous	17134
Squats (body weight)	Ambulatory/exercise	Moderate	02052
Biceps curls (1.4-kg dumbbell per hand)	Ambulatory/exercise	Light	02024
	Ball State University data		
Lying down	Sedentary	Sedentary	07010
Using a computer	Sedentary	Sedentary	07021
Watching television	Sedentary	Sedentary	07020
Writing	Sedentary	Sedentary	07050
Playing cards	Sedentary	Sedentary	07021
Reading quietly	Sedentary	Sedentary	07070
Standing	Household	Light	07041
Dusting	Household	Light	05032
Making bed	Household	Light	05040
Folding laundry	Household	Light	05090
Sweeping	Household	Light	05011
Vacuuming	Household	Light	05040
Gardening: scooping dirt with hand shovel	Household	Light	08135
Picking up items (<1 kg) off floor	Household	Moderate	05030
Walking slowly (overground)	Ambulatory/exercise	Light	17152
Walking briskly (overground)	Ambulatory/exercise	Moderate	17190
Self-paced walking (treadmill)	Ambulatory/exercise	Moderate	17190
Cycling (stationary, 75–150 W)	Ambulatory/exercise	Moderate	02017
Stair climbing/descending	Ambulatory/exercise	Vigorous	17134
Overground jogging	Ambulatory/exercise	Vigorous	12020
Treadmill jogging	Ambulatory/exercise	Vigorous	12020

For each interval, the following time-domain features were computed for each axis of the accelerometer: mean, standard deviation, minimum, maximum, and various percentiles (10th, 25th, 50th, 75th, and 90th); these features have been commonly utilized in past work (24, 38). To investigate frequency-domain features, the acceleration data over each interval were transformed with the fast Fourier transform. On the transformed data, the maximal amplitude and its corresponding frequency (called the "principal frequency") were recorded. Additionally, the sum of the amplitudes (called the "total integral") was computed, the partial sum of the amplitudes between 0.6 and 2.5 Hz (called the "integral between 0.6 and 2.5 Hz") was computed, and also the percentage of this integral over the total integral was recorded. Similar frequency-domain features have been used in past work (37, 45).

The feature sets that were investigated in this study can be found in Table 3. *Feature set 1* contains all of the time-domain features that were extracted, and it was chosen to be the reference set against which the other sets would be compared. *Feature sets 2–4* are subsets of *feature set 1*. *Feature sets 5* and *6* contain all the time-domain features and also contain frequency-domain features. *Feature set 5* contains all of the frequency-domain features that were extracted, and *feature set 6* contains only the principal frequency and peak amplitude.

Machine learning methods. Four machine learning methods were investigated, as well as majority voting approaches combining these methods. These methods were decision trees with boosting, random forests, artificial neural networks, and support vector machines.

A decision tree is a popular classification method (33). Each internal node in the tree consists of a binary split of an individual

Table 3. *Feature sets used in this study*

Feature Set	Features Used	Total No. of Features Used
1	Mean, SD, minimum, maximum, 10th, 25th, 50th, 75th, and 90th percentiles of acceleration signal	27 (9 features × 3 axes)
2	Minimum, maximum, 10th, 25th, 50th, 75th, and 90th percentiles of acceleration signal	21 (7 features × 3 axes)
3	Mean, SD, minimum and maximum of acceleration signal	12 (4 features × 3 axes)
4	Mean and SD of acceleration signal	6 (2 features × 3 axes)
5	Mean, SD, minimum, maximum, 10th, 25th, 50th, 75th, and 90th percentiles, principal frequency, peak amplitude, integral between 0.6 and 2.5 Hz and percentage of total integral between 0.6 and 2.5 Hz of acceleration signal	39 (13 features × 3 axes)
6	Mean, SD, minimum, maximum, 10th, 25th, 50th, 75th, and 90th percentiles, principal frequency and peak amplitude of acceleration signal	33 (11 features × 3 axes)

feature extracted from the raw data, and the predicted class for an observation in the test data is made when a leaf of the tree is reached. For this study, the C5.0 algorithm from the C50 package in R (R Core Development Team, Vienna, Austria), was used to train the trees. The classification performance of decision trees can be improved by the process of boosting (34). Boosting is an iterative procedure in which a sequence of classification algorithms is trained, each subsequent algorithm giving more weight to the observations that were classified incorrectly by the previous algorithm. In this study, 15 boosting iterations were used.

A random forest is a classification method that trains a large number of decision trees and uses the majority predicted class from these trees as the final prediction (17). When each decision tree is trained, only a random subset of the features is considered at each split in the tree, leading to a wide variety of trees. For this study, the randomForest package in R was used. Each random forest contained 500 trees, and the size of the random subset of features at each split was the square root of the total number of features.

Artificial neural networks are nonlinear functions of the input features that are popularly tested methods for activity classification (32). For this study, the nnet package in R was used to train the artificial neural networks. Each network contained a single hidden layer of 15 nodes and was trained for a maximum of 200 iterations. The results from each individual neural network can be quite variable. Therefore, a majority voting system was tested in addition to individual networks. In the majority voting system, 30 separate neural networks were trained, and the majority predicted class from these networks was used as the final prediction. In case of a tie, a class was chosen at random from the tied classes.

Support vector machines are another popular classification method (40). A support vector machine performs classification via a set of hyperplanes. Because classes are often not linearly separable in the original feature space, typically the features are projected by a "kernel" function into a higher-dimensional space, where a linear classifier may be more successful. For this study, the kernlab package in R was used to train the support vector machines, and the radial basis kernel function was used.

Finally, a majority voting method was implemented to combine the random forests, decision trees, neural networks (majority of 30), and support vector machines. Whichever predicted class was most common among the four methods was adopted as the final prediction for the majority method. In case of a tie, a class was chosen at random from the tied classes.

Cross-validation and out-of-sample testing. Leave-one-out cross-validation was used as an initial step in comparing accuracy of the feature sets and machine learning methods. In cross-validation, the original data are divided into several sections, called "folds" (16). A machine learning method is trained on all but one of the folds and used to predict the observations in the remaining fold. This procedure is repeated until all the folds have been predicted. In this study, the folds corresponded to the individual participants in each study ("leave-one-out" cross-validation). In other words, the activity intensities for each participant were predicted using the data from all the other participants. This procedure was performed separately for the MSU and BSU data sets.

A true out-of-sample test was also performed by training each classification method (with each feature set) on the entire MSU data set and predicting the BSU data set, and vice versa. The true out-of-sample test likely offers the best estimate of generalizability of the classification methods, and this test often yields lower accuracy than in initial development/validation (36). In the out-of-sample test, we also compared our methods to the euclidean norm minus one (ENMO) method, which has been investigated by other research groups to classify activity intensity using a threshold-based approach on raw data from wrist-worn accelerometers (3, 4). Mean ENMO signals of $<$30, 30–99, 100–399, and $\geq$400 milli-g were used to classify intervals as sedentary, light, moderate, or vigorous intensities, respectively. These are the default thresholds used in the GGIR package in the R statistical software. The ENMO/GGIR approach is described in more detail in past works (3, 4).

Data analysis. The predictive models were evaluated with classification accuracies. For each 30-s interval, the predicted activity intensity from each model was compared to the criterion-measured intensity and the percentage of correct classifications was calculated. To assess the uncertainty in the classification accuracies, approximate 95% bootstrap confidence intervals (CIs) were calculated (9). The bootstrap is a statistical method in which the original data are resampled, with replacement, to create many new data sets of the same size. In this study, the 2.5th and 97.5th percentiles of the bootstrap classification accuracies were used to form a CI (the "bootstrap percentile method"). Bootstrap CIs were calculated for a reference classification method and for the difference between the reference method and the other methods. An interval that did not contain 0 indicated a significant difference of a classification method from the reference method. The machine learning models, example data, and code to run the models are available online at https://drive.google.com/file/d/0B-BgdTzyd2OxMGlLR1ZhTjI0R28/view.

RESULTS

Leave-one-out cross-validation. Leave-one-out cross-validation was performed separately for the MSU and BSU data sets. Classification accuracies and approximate 95% bootstrap CIs for the six feature sets are given in Table 4, using a random forest as the classification method. The results from the random forest are presented because the random forest was determined in out-of-sample tests to be the most successful method (see Table 5 and Table 7). For *feature set 1* the CI is for the classification accuracy, while for *feature sets 2–6* the CI is for the difference between the classification accuracy for that feature set and *feature set 1*. Performance was better on the MSU data set than on the BSU data set. This same result was observed with different machine learning methods (see Table 5).

Feature sets 1 and *2* performed comparably. *Feature set 1* outperformed *feature set 3* on the MSU data set and outperformed *feature set 4* on both data sets. Therefore, the percentile features improved performance slightly compared with simpler feature sets. *Feature set 5* outperformed *feature set 1* on both data sets, and *feature set 6* outperformed *feature set 1* on the MSU data set. Therefore, classification accuracy was improved by including frequency-domain features.

Table 4. *Classification accuracies and bootstrap confidence intervals for each feature set with leave-one-out cross-validation*

Feature Set	Michigan State University Data Set	Ball State University Data Set
1	91.7 [90.7, 92.3]	78.4 [76.6, 80.5]
2	91.9 [−0.2, 0.5]	78.4 [−1.1, 0.5]
3	89.3 [−3.2, −1.7]*	79.0 [−1.3, 1.4]
4	87.8 [−4.9, −3.1]*	76.7 [−3.0, −0.1]*
5	92.8 [0.7, 1.6]*	80.2 [0.8, 3.1]*
6	92.6 [0.5, 1.4]*	79.3 [−0.2, 1.9]

Data (in %) are means [95% bootstrap confidence interval]. The classification method was a random forest. For *feature set 1*, the bootstrap confidence interval represents the 95% confidence interval surrounding the mean. For *feature sets 2–6*, the bootstrap confidence intervals represent the 95% confidence interval for the difference between a given feature set and *feature set 1*. *Significant difference from *feature set 1*.

Table 5. *Classification accuracies from out-of-sample testing for each feature set and machine learning method*

Out-of-Sample Testing	Feature Set 1	2	3	4	5	6
Train MSU, test BSU						
Random forest	77.3	77.1	76.6	75.5	76.1	77.0
Neural network (individual)	73.5	74.3	74.4	74.8	69.6	74.1
Neural network (majority of 30)	77.0	77.9	76.7	76.5	73.0	77.4
Decision tree	76.4	75.1	75.4	74.4	73.3	76.0
Support vector machine	76.1	76.1	74.4	74.7	69.4	76.1
Majority of methods	77.9	77.5	76.4	76.2	74.5	77.8
Train BSU, test MSU						
Random forest	78.5	79.1	76.2	74.1	74.0	77.4
Neural network (individual)	69.5	68.9	66.8	69.4	53.6	58.8
Neural network (majority of 30)	77.7	78.1	72.6	73.6	60.0	65.3
Decision tree	75.7	76.9	73.5	73.5	67.5	75.0
Support vector machine	70.9	70.5	69.4	71.6	61.3	68.1
Majority of methods	77.8	78.6	75.3	74.7	68.5	51.0

Data are in %. MSU, Michigan State University data set; BSU, Ball State University data set.

Out-of-sample tests. A true out-of-sample test was also performed for each of the feature sets, by training on the MSU data set and predicting the BSU data set, and vice versa. Classification accuracies with bootstrap CIs are given in Table 6, again using a random forest as the classification method.

Feature set 1 obtained an accuracy of 77.3% when training on MSU and testing on BSU, which was comparable to the 78.4% found in leave-one-out cross-validation for BSU, and an accuracy of 78.5% when training on BSU and testing on MSU, which was lower than the 91.7% found in the cross-validation for the MSU data in the leave-one-out cross-validation. This was generally true across all other feature sets, with classification accuracies no more than 4.1 percentage points lower between cross-validation and out-of-sample testing for the BSU data set but as much as 18.8 percentage points lower between cross-validation and out-of-sample testing for the MSU data set.

Feature sets 1 and *2* performed comparably on both data sets, *feature set 1* outperformed *feature set 3* when training on the BSU data set, and *feature set 1* outperformed *feature set 4* on both data sets. Therefore, the sets including percentiles were generally more successful than the sets without percentiles.

Table 6. *Classification accuracies and bootstrap confidence intervals for each feature set, training on one data set and testing on the other*

Feature Set	Train on MSU, Test on BSU	Train on BSU, Test on MSU
1	77.3 [74.9, 78.7]	78.5 [76.0, 79.9]
2	77.1 [−1.2, 0.6]	79.1 [−0.4, 1.5]
3	76.6 [−2.1, 0.6]	76.2 [−4.1, −0.5]*
4	75.5 [−3.0, −0.2]*	74.1 [−6.9, −2.6]*
5	76.1 [−2.2, 0.3]	74.0 [−7.1, −1.9]*
6	77.0 [−1.1, 0.8]	77.4 [−2.6, 0.5]

Data (in %) are shown as mean [95% bootstrap confidence interval]. The classification method was a random forest. For *feature set 1*, the bootstrap confidence interval represents the 95% confidence interval surrounding the mean. For *feature sets 2–6*, the bootstrap confidence intervals represent the 95% confidence interval for the difference between a given feature set and *feature set 1*. MSU, Michigan State University data set; BSU, Ball State University data set. *Significant difference from *feature set 1*.

Table 7. *Classification accuracies and bootstrap confidence intervals for each machine learning method, training on one data set and testing on the other*

Modeling Method	Train on MSU, Test on BSU	Train on BSU, Test on MSU
Random forest	77.3 [74.9, 78.7]	78.5 [76.0, 79.9]
Neural network (individual)	73.5 [−7.8, −1.3]*	69.5 [−15.8, −6.8]*
Neural network (majority of 30)	77.0 [−1.4, 1.7]	77.7 [−3.8, 0.7]
Decision tree	76.4 [−2.2, 1.5]	75.7 [−4.9, −0.3]*
Support vector machine	76.1 [−2.8, 0.5]	70.9 [−10.9, −5.5]*
Majority of methods	77.9 [−0.1, 1.8]	77.8 [−2.1, 0.6]
ENMO/GGIR	70.6 [−7.6, −3.6]*	53.6 [−25.2, −20.8]*

Data (in %) are shown as mean [95% bootstrap confidence interval]. *Feature set 1* was used. For random forest, the bootstrap confidence interval represents the 95% confidence interval surrounding the mean. For all machine learning methods except random forest, the bootstrap confidence intervals represent the 95% confidence interval for the difference between a given machine learning method and random forest. MSU, Michigan State University data set; BSU, Ball State University data set; ENMO/GGIR, the euclidean norm minus one method using the GGIR package in R; Neural network (majority of 30), the most commonly chosen activity intensity from 30 different neural networks; Majority of methods, the most commonly chosen activity intensity from random forest, neural network (majority of 30), decision tree, and support vector machine models. *Significant difference from random forest.

The same result was observed in cross-validation. *Feature set 1* performed comparably to *feature set 6* on both data sets and outperformed *feature set 5* when training on the BSU data set. Therefore, the frequency-domain features did not improve performance, and even harmed accuracy in some cases. This differs from the result of cross-validation, where the frequency-domain features improved accuracy.

Comparison of machine learning methods and ENMO/GGIR. Out-of-sample tests were also performed to compare the machine learning methods. The methods were trained on the MSU data set and used to predict the BSU data set, and vice versa. Classification accuracies for *feature set 1* are given in Table 7, which also includes approximate 95% bootstrap CIs for the random forest and the differences between the other classifiers and the random forest. Similar results were observed for different feature sets (classification accuracies are shown in table).

The random forests, neural networks (majority of 30), decision trees, and support vector machines performed comparably when training on the MSU data set and testing on the BSU data set. When training on the BSU data set, the random forests and neural networks (majority of 30) performed comparably and outperformed the decision trees and support vector machines. The individual neural networks performed significantly worse on both data sets. The majority voting method did not significantly improve performance on either data set. The ENMO/GGIR method performed significantly worse than the random forest on both data sets, but the difference was more pronounced when training on the BSU data set and testing on the MSU data set. Since the random forests performed comparably to the neural networks (majority of 30) and outperformed the other methods in some cases, but are simpler to train and use than the majority neural networks, results from random forest models were focused on for presentation in this report (Tables 4, 6, 8, and 9).

Comparison of age groups. The BSU data set included a much larger range of participant ages than the MSU data set.

Table 8. *Classification accuracies and bootstrap confidence intervals for the three age groups, training on Michigan State University data set and predicting on Ball State University data set*

Age Group	Accuracy, %
Under 40	77.5 [74.5, 80.3]
40–60	75.4 [−5.7, 2.0]
Over 60	79.8 [−4.1, 4.9]

Data are shown as mean [95% bootstrap confidence interval]. *Feature set 1* was used. The classification method was a random forest. For Under 40, the bootstrap confidence interval represents the 95% confidence interval surrounding the mean. For 40–60 and Over 60, the bootstrap confidence intervals represent the 95% confidence interval for the difference between a given feature set and Under 40. Analyses revealed no statistically significant different differences in accuracy among groups.

Therefore, we investigated whether classification performance was different for the ages not present in the MSU data set when training on the MSU data set and predicting the BSU data set. The participants in the BSU data set were divided into three age groups: Under 40 yr old, 40–60 yr old, and Over 60 yr old. Classification accuracies for each age group are given in Table 8, using *feature set 1* and a random forest as the classification method. Table 8 also includes approximate 95% bootstrap CIs for the first age group (Under 40) and the difference between the other age groups and the first age group. Classification accuracy was comparable for the Under 40 group and the two older groups. Similar results were observed for different machine learning methods and feature sets (data not shown).

Prediction accuracy by activity type. The activity-specific prediction accuracies were also calculated. The prediction accuracy for a particular activity refers to the percentage of all 30-s intervals of that activity that were correctly predicted to be the intensity corresponding to that activity. Results for the random forest classifier with *feature set 1* are given in Table 9. Classification accuracy was >80% for all sedentary activities except watching television (BSU data set). Similarly, classification accuracy was >80% for all light-intensity activities except walking slowly (MSU data set), which was generally misclassified as moderate intensity, and gardening (BSU data set). In contrast, many moderate- and vigorous-intensity activities had classification accuracies <80%. In both data sets, cycling was often misclassified as sedentary or light intensity and stair climbing/descending was often misclassified as light or moderate intensity. In the MSU data set squats was almost always misclassified as sedentary activity, while in the BSU data set picking up items was almost always misclassified as light intensity. In the BSU data set, brisk and treadmill walking were often misclassified as light intensity and both overground and treadmill jogging were often misclassified as light or moderate intensity.

DISCUSSION

Our results indicate that a random forest model coupled with time-domain features attained high accuracy for activity intensity prediction from a wrist-worn accelerometer in both cross-validation and out-of-sample testing on an independent population. Additionally, the machine learning models developed and tested in this study had higher accuracy for activity intensity prediction than the ENMO/GGIR method, which is arguably the most common current method of determining activity intensity from wrist-worn accelerometers (3, 4, 35). Wrist motion does not necessarily correlate well with energy expenditure or activity intensity (37, 41), so it is not surprising that ENMO/GGIR, which determines activity intensity solely on thresholds of the vector magnitude of the raw signal, had lower accuracy than the machine learning models. That the machine learning models outperformed ENMO/GGIR even in independent out-of-sample testing suggests that they have better generalizability and should be preferred over ENMO/GGIR when classifying activity intensity with wrist-worn accelerometers. As wrist-worn accelerometers continue to gain popularity because of improved comfort, compliance, and ability to measure constructs such as activity type and sleep, it is apparent that machine learning models will be instrumental in obtaining accurate PA metrics from these data (14, 20, 42).

Our study revealed a number of important considerations when using machine learning to predict PA variables from wrist-worn accelerometers. First, our study suggests that leave-one-out cross-validation may not be a sufficient method to determine accuracy/utility of machine learning models and/or feature sets for predicting PA metrics. In leave-one-out cross-validation, classification accuracies for the random forest were 10.3–13.5 percentage points higher when conducted within the

Table 9. *Prediction accuracies for each activity type*

Activity Type	Activity Intensity	Prediction Accuracy, %
Training on MSU, predicting BSU		
Lying down	Sedentary	86.8
Using a computer	Sedentary	89.3
Watching television	Sedentary	70.3
Writing	Sedentary	100.0
Playing cards	Sedentary	97.9
Reading quietly	Sedentary	87.2
Standing	Light	85.3
Dusting	Light	88.2
Making bed	Light	84.8
Folding laundry	Light	90.0
Sweeping	Light	88.6
Vacuuming	Light	85.7
Gardening	Light	76.9
Picking up items	Moderate	9.1
Walking slowly	Light	88.2
Walking briskly	Moderate	68.4
Self-paced walking (treadmill)	Moderate	38.6
Cycling	Moderate	68.2
Stair climbing/descending	Vigorous	54.1
Overground jogging	Vigorous	77.4
Treadmill jogging	Vigorous	54.2
Training on BSU, predicting MSU		
Lying down	Sedentary	98.5
Reading quietly	Sedentary	93.7
Using a computer	Sedentary	95.2
Standing	Light	95.9
Sweeping	Light	86.9
Folding laundry	Light	93.9
Walking slowly	Light	31.8
Walking briskly	Moderate	89.6
Jogging	Vigorous	99.8
Cycling	Moderate	74.7
Stair climbing/descending	Vigorous	61.0
Squats	Moderate	2.9
Biceps curls	Light	80.9

Feature set 1 was used. The classification method was a random forest. MSU, Michigan State University data set; BSU, Ball State University data set.

MSU data set compared with the BSU data set. However, out-of-sample testing yielded similar classification accuracies between data sets (74.0–78.5% for MSU, 75.5–77.3% for BSU), which were much closer to the accuracies obtained within the BSU leave-one-out cross-validation. The MSU data set had fewer activities and less interparticipant variability in age, body composition, etc., than the BSU data set; it appears that the increased variability of the BSU data set gave the leave-one-out cross-validation a closer representation of accuracy that could be expected in true out-of-sample testing. Nevertheless, leave-one-out cross-validation within the BSU data set indicated that frequency-domain features improved predictive accuracy, but this did not occur in out-of-sample testing. The results of the present study are supported by a study by Montoye et al. in which machine learning models were developed for energy expenditure prediction in children (A. H. Montoye, K. A. Pfeiffer, K. A. Clevenger, K. A. Mackintosh, and M. A. McNarry, unpublished observations). In leave-one-out cross-validation, there was minimal difference in accuracy among accelerometer placements (hip, wrist, or combination of hip and wrist) or data type (raw or count based) used in the machine learning models. However, out-of-sample testing in a data set with different activities and a slightly older population of children revealed significant differences in classification accuracies and differences in performance of the modeling techniques. The disagreement in accuracy when models are tested in different settings suggests that the leave-one-out approach is not sufficient on its own for understanding how developed models will perform in a new population. In other words, assessment of predictive models should be conducted in several data sets and/or settings to obtain a better idea of *1*) the expected accuracy of a given model and *2*) the comparative accuracy of machine learning models, feature sets, accelerometer placements, etc., when trying to identify optimal methods for free-living ("real world") PA assessment.

Second, the random forests achieved the best accuracy of the machine learning models in out-of-sample testing, with little difference in accuracy among the other nonensemble machine learning models. Specifically, the random forests outperformed the individual neural networks by 3.8 and 9.0 percentage points on the two data sets, outperformed the decision trees by 2.8 percentage points and the support vector machines by 7.6 percentage points when using MSU as the test data set, and performed comparably to the decision trees and support vector machines when using BSU as the test data set. The majority voting methods performed comparably to the random forests, but these methods are more complicated to train and use. Also, the majority of 30 neural networks outperformed the individual neural networks by 3.5 and 8.2 percentage points on the two data sets, suggesting that researchers should be cautious when using individual neural networks for predicting activity intensity and should consider comparing them with other approaches.

In related work, Chowdhury et al. (6) also found that random forests achieved the highest accuracies of seven conventional machine learning models tested on three data sets. However, Chowdhury et al. also found that custom-built ensemble methods outperformed the random forests, which differs from our results. Zhang et al. (45) and Dong et al. (8) found that various machine learning methods all performed similarly (differences of no more than 5 percentage points), although random forests were not investigated in these studies. All three of these previous studies used cross-validation rather than true out-of-sample testing. More studies are needed to determine whether random forests maintain an advantage over other machine learning models in a variety of settings. It also needs to be established whether the observed differences between models are clinically/practically significant.

Third, our study found no benefit to including frequency-domain features in the machine learning models. Feature sets including frequency-domain features had slightly higher accuracy in leave-one-out cross-validation; however, they did not improve accuracy in out-of-sample testing and in some cases resulted in significantly lower accuracy. This suggests that frequency-domain features may increase the risk of overfitting a machine learning model to the training data set, resulting in high initial accuracy but poor generalizability. These findings lie in contrast to results from Ellis et al. (10), who developed random forest models to predict four activity types in a free-living setting. They found that frequency-domain features aided significantly in predictive accuracy with a wrist-worn accelerometer but less so with a hip-worn accelerometer. Previous findings by Montoye et al. (22) also indicate that for energy expenditure prediction simpler feature sets work well for hip-worn accelerometers, but more complex feature sets were needed to achieve high accuracy from wrist-worn accelerometers. However, in both of these previous studies, the machine learning models were evaluated with leave-one-out cross-validation, which may not yield results equivalent to true out-of-sample testing (as previously discussed). Further testing is needed to identify feature sets that provide a balance of complexity/ease of use and accuracy across a variety of settings and data sets, noting also that the optimal feature set is likely different depending on accelerometer placement.

Finally, we note that our machine learning models had higher accuracy for classifying sedentary and light-intensity activities compared with moderate- or vigorous-intensity activities. The only sedentary activity frequently misclassified was watching television, likely because of movement of the wrist while changing television channels with a remote control. We surmise that moderate- and vigorous-intensity activities were misclassified more often because of the increased variability in motion when performing higher-intensity activities, especially in the BSU data set. For example, the jogging speed of an older participant was often similar to the brisk walking speed of a younger participant; this caused the models to have difficulty distinguishing these activities. Similarly, there was a considerable range of speeds at which individuals walked as well as climbed and descended stairs. This is reflected in the poor classification accuracy for stair climbing/descending and for the leisure, brisk, and self-paced walking activities. We considered but ultimately rejected several options to try to further improve classification of moderate- and vigorous-intensity activities. First, we considered grouping all walking activities into the moderate-intensity category to improve accuracy but decided against this since the Compendium of Physical Activities and United States PA recommendations suggest that not all walking is of sufficient intensity to be health enhancing (i.e., of at least moderate intensity) (1, 31). We also considered collapsing moderate and vigorous intensities into a single "MVPA" category, which has

been done in past work (28); such an approach would not help with the light-moderate misclassifications but would alleviate the moderate-vigorous misclassifications. However, this approach would make determination of meeting PA guidelines more difficult since individuals need to perform less MVPA if intensity is vigorous than if it is moderate.

Given the considerable speed variability in moderate- and vigorous-intensity activities performed by our study participants, it is not surprising that these were most difficult to classify with our models developed from the BSU data set. We prioritized achieving reasonably accurate models across a varied sample so that our models had better generalizability across groups of varying ages and fitness levels. We were successful in this regard, with no significant differences observed in classification accuracy when comparing across age groups (Table 8). In other words, the variability in performance of higher-intensity activities did not lead to particularly poor classification accuracy for older participants but instead lowered accuracy for all participants. Studies seeking to optimize accuracy of classifying moderate- and vigorous-intensity activities for a specific group may elect to use models similar to our models from the MSU data set, developed with the leave-one-out approach (developed in healthy, younger adults), since those were able to achieve much higher accuracy than the models from the BSU data set (a much more varied sample). An alternative route to try to improve classification of moderate- and vigorous-intensity activities would be to collect more data specifically with activities in these intensities, which would allow more training data, to potentially improve the ability of the models to differentiate between these intensity categories. It may be especially helpful to include activities at the high end of light intensity, the low end of moderate intensity, the high end of moderate intensity, and the low end of vigorous intensity; while such an approach would require addition of data from a stricter laboratory protocol to ensure that these intensities are captured, the advantage of this approach is that it may allow the models to better determine the boundaries among intensity levels. These possibilities should be explored in future work.

This study has several notable strengths. First, this study evaluated several machine learning models and majority voting approaches across a wide range of feature sets, giving good representation of different methodological approaches for achieving optimal activity intensity prediction with a wrist-worn accelerometer. Second, the inclusion of two data sets incorporating a diverse population and many distinct activities allowed for two types of machine learning training and testing, giving this study strong potential generalizability to other populations and settings. Third, we view the use of direct observation and automatic activity intensity coding as a study strength; Lyden et al. (19) demonstrate that this method is translatable to free-living settings, and recent studies suggest that direct observation or similar methods (e.g., body-worn camera) is becoming an accepted criterion measure for free-living PA assessment (10, 13).

Our study also has several limitations. First, the sample size of each of our individual data sets is relatively small, and the MSU data set had little variability in age. Second, while both of these data sets were collected in simulated free-living protocols meant to increase generalizability of study findings, the studies ultimately took place within a laboratory setting, which reduces variability in the types of activities the participants could perform and the manner in which they could perform activities. The developed machine learning models should be validated in a true free-living setting in order to confirm this study's findings. Third, our use of direct observation resulted in arbitrary choices that had to be made regarding the intensity of activities performed. For example, the Compendium of Physical Activities (1) allows cycling to be coded as light, moderate, or vigorous intensity depending on the resistance the participant chose in the protocols and the participant's body weight. We elected to code activities by type and automatically assign all activities of the same type the same intensity value (e.g., all cycling activities were coded as moderate intensity). This choice was meant to increase objectivity of direct observation instead of asking observers to subjectively assess the intensity of the task; we felt this would have led to more errors within the observation because of the large number of tasks and the variability in which they were performed. Finally, by necessity we only included windows of data with a single activity performed and excluded transitions and windows with multiple activities. Transitions are notoriously difficult to classify, especially for energy expenditure, and our models may have reduced accuracy when attempting to classify free-living data containing transitions between activities (2).

In conclusion, machine learning models developed to predict activity intensity from an accelerometer worn on the left wrist achieved high accuracy in both leave-one-out cross-validation and out-of-sample testing. However, results of out-of-sample testing differed from results of leave-one-out (within sample) cross-validation, suggesting that leave-one-out cross-validation is not a sufficient method to evaluate machine learning models. The random forest was the most consistently high-performing machine learning model. No apparent benefit was found when using majority voting approaches or frequency-domain features, meaning that a single machine learning model with time-domain features may provide an optimal blend of accuracy and model complexity for predicting activity intensity from a wrist-worn accelerometer.

ACKNOWLEDGMENTS

The study authors thank Ball State University and Michigan State University for sharing the de-identified data sets for use in this study. A. H. K. Montoye is principal investigator on the BSU data set, and K. A. Pfeiffer is principal investigator on the MSU data set.

This article reflects the viewpoints of the study authors only.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

ENDNOTE

At the request of the authors, readers are herein alerted to the fact that additional materials related to this manuscript may be found at the institutional Web site of the authors, which at the time of publication they indicate is: https://drive.google.com/file/d/0B-BgdTzyd2OxMGlLR1ZhTjI0R28/view. These materials are not a part of this manuscript and have not undergone peer review by the American Physiological Society (APS). APS and the journal editors take no responsibility for these materials, for the Web site address, or for any links to or from it.

AUTHOR CONTRIBUTIONS

A.H.M., B.S.W., M.R.F., and K.A.P. conceived and designed research; A.H.M. performed experiments; A.H.M. and B.S.W. analyzed data; A.H.M.,

B.S.W., M.R.F., and K.A.P. interpreted results of experiments; B.S.W. prepared figures; A.H.M. and B.S.W. drafted manuscript; A.H.M., B.S.W., M.R.F., and K.A.P. edited and revised manuscript; A.H.M., B.S.W., M.R.F., and K.A.P. approved final version of manuscript.

REFERENCES

1. **Ainsworth BE, Haskell WL, Herrmann SD, Meckes N, Bassett DR Jr, Tudor-Locke C, Greer JL, Vezina J, Whitt-Glover MC, Leon AS.** 2011 Compendium of Physical Activities: a second update of codes and MET values. *Med Sci Sports Exerc* 43: 1575–1581, 2011. doi:10.1249/MSS.0b013e31821ece12.
2. **Altini M, Penders J, Amft O.** Estimating oxygen uptake during non-steady-state activities and transitions using wearable sensors. *IEEE J Biomed Health Inform* 20: 469–475, 2016. doi:10.1109/JBHI.2015.2390493.
3. **Bai J, Di C, Xiao L, Evenson KR, LaCroix AZ, Crainiceanu CM, Buchner DM.** An activity index for raw accelerometry data and its comparison with other activity metrics. *PLoS One* 11: e0160644, 2016. doi:10.1371/journal.pone.0160644.
4. **Bakrania K, Yates T, Rowlands AV, Esliger DW, Bunnewell S, Sanders J, Davies M, Khunti K, Edwardson CL.** Intensity thresholds on raw acceleration data: euclidean norm minus one (ENMO) and mean amplitude deviation (MAD) approaches. *PLoS One* 11: e0164045, 2016. doi:10.1371/journal.pone.0164045.
5. **Bastian T, Maire A, Dugas J, Ataya A, Villars C, Gris F, Perrin E, Caritu Y, Doron M, Blanc S, Jallon P, Simon C.** Automatic identification of physical activity types and sedentary behaviors from triaxial accelerometer: laboratory-based calibrations are not enough. *J Appl Physiol (1985)* 118: 716–722, 2015. doi:10.1152/japplphysiol.01189.2013.
6. **Chowdhury AK, Tjondronegoro D, Chandran V, Trost SG.** Ensemble methods for classification of physical activities from wrist accelerometry. *Med Sci Sports Exerc* 49: 1965–1973, 2017. doi:10.1249/MSS.0000000000001291.
7. **Donaldson SC, Montoye AH, Tuttle MS, Kaminsky LA.** Variability of objectively measured sedentary behavior. *Med Sci Sports Exerc* 48: 755–761, 2016. doi:10.1249/MSS.0000000000000828.
8. **Dong B, Montoye A, Moore R, Pfeiffer K, Biswas S.** Energy-aware activity classification using wearable sensor networks. *Proc SPIE* 8723: 87230Y, 2013. doi:10.1117/12.2018134.
9. **Efron B.** Bootstrap methods: another look at the jackknife. *Ann Stat* 7: 1–26, 1979. doi:10.1214/aos/1176344552.
10. **Ellis K, Kerr J, Godbole S, Staudenmayer J, Lanckriet G.** Hip and wrist accelerometer algorithms for free-living behavior classification. *Med Sci Sports Exerc* 48: 933–940, 2016. doi:10.1249/MSS.0000000000000840.
11. **Freedson PS, Melanson E, Sirard J.** Calibration of the Computer Science and Applications, Inc. accelerometer. *Med Sci Sports Exerc* 30: 777–781, 1998. doi:10.1097/00005768-199805000-00021.
12. **Gyllensten IC, Bonomi AG.** Identifying types of physical activity with a single accelerometer: evaluating laboratory-trained algorithms in daily life. *IEEE Trans Biomed Eng* 58: 2656–2663, 2011. doi:10.1109/TBME.2011.2160723.
13. **Hickey A, Del Din S, Rochester L, Godfrey A.** Detecting free-living steps and walking bouts: validating an algorithm for macro gait analysis. *Physiol Meas* 38: N1–N15, 2017. doi:10.1088/1361-6579/38/1/N1.
14. **Jean-Louis G, Kripke DF, Cole RJ, Assmus JD, Langer RD.** Sleep detection with an accelerometer actigraph: comparisons with polysomnography. *Physiol Behav* 72: 21–28, 2001. doi:10.1016/S0031-9384(00)00355-3.
15. **Kerr J, Patterson RE, Ellis K, Godbole S, Johnson E, Lanckriet G, Staudenmayer J.** Objective assessment of physical activity: classifiers for public health. *Med Sci Sports Exerc* 48: 951–957, 2016. doi:10.1249/MSS.0000000000000841.
16. **Kohavi R.** A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc Conf AAAI Artif Intell* 14: 1137–1145, 1995.
17. **Liaw A, Wiener M.** Classification and regression by randomForest. *R News* 2: 18–22, 2002.
18. **Lyden K, Keadle SK, Staudenmayer J, Freedson PS.** A method to estimate free-living active and sedentary behavior from an accelerometer. *Med Sci Sports Exerc* 46: 386–397, 2014. doi:10.1249/MSS.0b013e3182a42a2d.
19. **Lyden K, Petruski N, Staudenmayer J, Freedson P.** Direct observation is a valid criterion for estimating physical activity and sedentary behavior. *J Phys Act Health* 11: 860–863, 2014. doi:10.1123/jpah.2012-0290.
20. **Mannini A, Intille SS, Rosenberger M, Sabatini AM, Haskell W.** Activity recognition using a single accelerometer placed at the wrist or ankle. *Med Sci Sports Exerc* 45: 2193–2203, 2013. doi:10.1249/MSS.0b013e31829736d6.
21. **Matthews CE, Chen KY, Freedson PS, Buchowski MS, Beech BM, Pate RR, Troiano RP.** Amount of time spent in sedentary behaviors in the United States, 2003–2004. *Am J Epidemiol* 167: 875–881, 2008. doi:10.1093/aje/kwm390.
22. **Montoye AH, Begum M, Henning Z, Pfeiffer KA.** Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data. *Physiol Meas* 38: 343–357, 2017. doi:10.1088/1361-6579/38/2/343.
23. **Montoye AH, Moore RW, Bowles HR, Korycinski R, Pfeiffer KA.** Reporting accelerometer methods in physical activity intervention studies: a systematic review and recommendations for authors. *Br J Sports Med* 2016: bjsports-2015-095947, 2016. doi:10.1136/bjsports-2015-095947.
24. **Montoye AH, Mudd LM, Biswas S, Pfeiffer KA.** Energy expenditure prediction using raw accelerometer data in simulated free living. *Med Sci Sports Exerc* 47: 1735–1746, 2015. doi:10.1249/MSS.0000000000000597.
27. **Montoye AH, Pivarnik JM, Mudd LM, Biswas S, Pfeiffer KA.** Comparison of activity type classification accuracy from accelerometers worn on the wrists, hip, and thigh in young, apparently healthy adults. *Meas Phys Educ Exerc Sci* 20: 173–183, 2016. doi:10.1080/1091367X.2016.1192038.
28. **Montoye AH, Pivarnik JM, Mudd LM, Biswas S, Pfeiffer KA.** Validation and comparison of accelerometers worn on the hip, thigh, and wrists for measuring physical activity and sedentary behavior. *AIMS Public Health* 3: 298–312, 2016. doi:10.3934/publichealth.2016.2.298.
29. **Ozemek C, Cochran HL, Strath SJ, Byun W, Kaminsky LA.** Estimating relative intensity using individualized accelerometer cutpoints: the importance of fitness level. *BMC Med Res Methodol* 13: 53, 2013. doi:10.1186/1471-2288-13-53.
30. **Paul DR, Kramer M, Moshfegh AJ, Baer DJ, Rumpler WV.** Comparison of two different physical activity monitors. *BMC Med Res Methodol* 7: 26, 2007. doi:10.1186/1471-2288-7-26.
31. **Physical Activity Guidelines Advisory Committee.** *2008 Physical Activity Guidelines for Americans*. Rockville, MD: Office of Disease Prevention and Health Promotion, 2008.
32. **Preece SJ, Goulermas JY, Kenney LP, Howard D, Meijer K, Crompton R.** Activity identification using body-mounted sensors—a review of classification techniques. *Physiol Meas* 30: R1–R33, 2009. doi:10.1088/0967-3334/30/4/R01.
33. **Quinlan JR.** *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann, 1993.
34. **Quinlan JR.** Bagging, boosting, and C4.5. *Proc Conf AAAI Artif Intell* 1: 725–730, 1996.
35. **Rowlands AV, Yates T, Davies M, Khunti K, Edwardson CL.** Raw accelerometer data analysis with GGIR R-package: does accelerometer brand matter? *Med Sci Sports Exerc* 48: 1935–1941, 2016. doi:10.1249/MSS.0000000000000978.
36. **Sasaki JE, Hickey AM, Staudenmayer JW, John D, Kent JA, Freedson PS.** Performance of activity classification algorithms in free-living older adults. *Med Sci Sports Exerc* 48: 941–950, 2016. doi:10.1249/MSS.0000000000000844.
37. **Staudenmayer J, He S, Hickey A, Sasaki J, Freedson P.** Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. *J Appl Physiol (1985)* 119: 396–403, 2015. doi:10.1152/japplphysiol.00026.2015.
38. **Staudenmayer J, Pober D, Crouter S, Bassett D, Freedson P.** An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *J Appl Physiol (1985)* 107: 1300–1307, 2009. doi:10.1152/japplphysiol.00465.2009.
39. **Strath SJ, Bassett DR Jr, Swartz AM.** Comparison of MTI accelerometer cut-points for predicting time spent in physical activity. *Int J Sports Med* 24: 298–303, 2003. doi:10.1055/s-2003-39504.
40. **Suykens JA, Vandewalle J.** Least squares support vector machine classifiers. *Neural Process Lett* 9: 293–300, 1999. doi:10.1023/A:1018628609742.
41. **Swartz AM, Strath SJ, Bassett DR Jr, O'Brien WL, King GA, Ainsworth BE.** Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. *Med Sci Sports Exerc* 32, *Suppl*: S450–S456, 2000. doi:10.1097/00005768-200009001-00003.

42. **Troiano RP, McClain JJ, Brychta RJ, Chen KY.** Evolution of accelerometer methods for physical activity research. *Br J Sports Med* 48: 1019–1023, 2014. doi:10.1136/bjsports-2014-093546.
43. **Trost SG, Loprinzi PD, Moore R, Pfeiffer KA.** Comparison of accelerometer cut points for predicting activity intensity in youth. *Med Sci Sports Exerc* 43: 1360–1368, 2011. doi:10.1249/MSS.0b013e318206476e.
44. **Welk GJ.** Use of accelerometry-based activity monitors to assess physical activity. In: *Physical Activity Assessments for Health-Related Research*, edited by Welk GJ. Champaign, IL: Human Kinetics, 2002, p. 125–142.
45. **Zhang S, Rowlands AV, Murray P, Hurst TL.** Physical activity classification using the GENEA wrist-worn accelerometer. *Med Sci Sports Exerc* 44: 742–748, 2012. doi:10.1249/MSS.0b013e31823bf95c.