Journal of Biomedical Analytics
Vol. 3, No. 1 (2020), pp. 27–50
doi:10.30577/jba.v3i1.57

Research Article

Statistical Learning Methods to Predict Activity Intensity from Body-Worn Accelerometers

Drew M Lazar[1,*], Munni Begum[1], Md Monzur Murshed[1], Benjamin Nelson[2], Joshua M Bock[2], Mary Imboden[2], Leonard Kaminsky[2], and Alex HK Montoye[3]

[1]Department of Mathematical Sciences, Ball State University, USA
[3]Integrative Physiology and Health Science, Alma College, USA
[*]Corrresponding author: dmlazar@bsu.edu

Received: June 20, 2020; revised: August 10, 2020; accepted: August 16, 2020.

Abstract: Physical activity, especially when performed at moderate or vigorous intensity, has short- and long-term health benefits, but measurement of free-living physical activity is challenging. Accelerometers are popular tools to assess physical activity, although accuracy of conventional accelerometer analysis methods is suboptimal. This study developed and tested statistical learning models for assessing activity intensity from body-worn accelerometers. Twenty-eight adults performed 10-21 activities of daily living in two visits while wearing four accelerometers (right hip, right ankle, both wrists). Accelerometer placement is of crucial practical concern and this paper addresses this issue. Boosting, bagging, random forest and decision tree models were created for each accelerometer and for two-, three-, and four-accelerometer combinations to predict activity intensity. Research staff observations of activity intensity served as the criterion. Point estimates of error for the ankle accelerometer were 2.2-4.7 percentage points lower than other single-accelerometer placements, and the left wrist-ankle combination had errors 0.8-5.8 percentage points lower than other two-accelerometer combinations. Decision trees had poorer accuracy than the other models. Using an accelerometer worn on the lower limb, by itself or in combination with an upper-limb accelerometer, appears to offer optimal accuracy for activity intensity measurement.

Keywords: accelerometers, physiology, machine learning, classification, regression trees, C5.0

1 Introduction

Participation in moderate or vigorous intensity physical activity (MVPA) has long been recognized as an important, modifiable risk factor in the development of cardiovascular and a host of other chronic diseases in adults of all ages (Pate *et al.*, 1995). More recently, there has been an emergence of evidence implicating sedentary behavior (SB), defined as seated or lying activities requiring low energy expenditure, as a possible independent, modifiable risk factor for health (Ekelund *et al.*, 2016). Given the mounting evidence that both MVPA and SB influence health, in 2018 the US Department of Health and Human Services updated its original 2008 guidelines, for the first time addressing participation in both MVPA and SB (Piercy *et al.*, 2018). However, the guidelines for SB do not give specific time recommendations, focusing instead on reducing SB and replacing it with light-, moderate-, or vigorous-intensity activity. Additionally, it is speculated that participation in MVPA may reduce or eliminate the chronic disease risks associated with spending high amounts of time spent in SBs, although this is still not fully understood or accepted (Ekelund *et al.*, 2016). Contributing to the lack of consensus on the level of contribution of MVPA and SB to health is the shortage of accurate, objective tools capable of measuring both of these behaviors in free-living settings.

In order to determine which activities contribute toward MVPA and SB recommendations, activities are typically divided into intensity levels (sedentary, light, moderate, vigorous) according to body position and energy expenditure required for the task. Seated or lying activities requiring ≤ 1.5 times the resting energy expenditure level (i.e., 1.5 metabolic equivalents or METS) are defined as SBs (Tremblay *et al.*, 2017), and activities eliciting 1.6-2.9, 3.0-5.9, and ≥ 6.0 METs are light, moderate, and vigorous intensities, respectively. Therefore, when assessing MVPA and SB, it is possible to do so by 1) first assessing the energy expenditure required of activities (without determining the types of activities performed) and translating these to activity intensities, 2) assessing the activity types being performed and using a tool such as the Compendium of Physical Activities (Ainsworth *et al.*, 2011) to assign energy expenditures and intensities to each activity type, or 3) assessing activity intensity directly without determining energy cost or activity types performed.

Accelerometer-based monitoring devices (hereafter referred to as accelerometers) are increasingly used in large scale studies as a means of measuring MVPA and SB and have recently been recommended as the preferred method for assessment of physical activity in populations such as those undergoing cardiac rehabilitation (Kaminsky *et al.*, 2016). Traditionally, accelerometers were worn on the hip to capture vertical movement of the trunk. Due to battery life and memory limitations of early accelerometers, proprietary and manufacturer-specific "activity counts" were derived from the accelerometer data in 1-60 second intervals (epochs), and cut-points were developed to determine the intensity of activities being performed in a given interval (John and Freedson, 2012). While activity counts and cut-points showed high accuracy for assessing physical activity intensity for hip-worn monitors during ambulatory activities (Freedson *et al.*, 1998), the relationship between activity counts and activity intensity is poorly defined for non-ambulatory activities, resulting in cut-points being relevant only to the population and types of activities for which the cut-points were validated. Thus, for development of modeling approaches that work well in a free-living context, inclusion of a variety of ambulatory and non-ambulatory activities appears necessary.

Technological improvements of accelerometers (device miniaturization, battery/memory improvements) have been accompanied by more advanced analytic methods for determining physical activity and SB. Monitors worn on alternate body locations have allowed for increased compliance (e.g., wrist; Troiano *et al.* (2014)) and accuracy (e.g., ankle placement for measuring steps; (Toth *et al.*, 2018)), and raw data capture has allowed for improved assessment of physical activity and SB. A recent review by Farrahi *et al.* (2019) found that most modeling methods for raw accelerometer data have been developed to assess either energy expenditure or activity type to then determine time spent in physical activity intensities. However, it is also possible to bypass these and assess activity intensity directly by treating each activity intensity as if it is a distinct activity type.

There are two advantages to assessing activity intensity directly rather than assessing energy expenditure or activity type first and then correspondingly deriving activity intensity. First, most modeling methods which assess energy expenditure have high error rates (e.g., root mean square error of 1-2 METs) and show "bias toward the mean", where the developed models overestimate the energy cost of low-intensity activities and underestimate the energy cost of high-intensity activities (Montoye *et al.*, 2015; Staudenmayer *et al.*, 2015). This generally leads to underestimation of time spent in SBs and vigorous-intensity activities and overestimation of time spent in light- and moderate-intensity activities. Second, prediction of activity type followed by activity intensity is problematic because it is not possible to develop a model to correctly classify all activity types in a free-living setting. Additionally, attempts made by recent work to classify activity types into broad categories based on similar activity patterns (Sasaki *et al.*, 2016; Kerr *et al.*, 2016) result in collapsing activities of different intensities into the same category, which is problematic for assessing adherence to physical activity recommendations.

Despite it being much more common to assess energy expenditure or activity type, there are several recent studies which have used accelerometers and machine learning to assess activity intensity directly. In one, we developed artificial neural network machine learning models to predict activity intensity as a three-class variable (sedentary, light, MVPA) from accelerometers located on the wrists, thigh, and hip (Montoye *et al.*, 2016). Overall, the thigh accelerometer placement was found to have superior accuracy to other placements, with the left wrist providing the second highest accuracy. Limitations to this study were that only one type of machine learning model and one set of input features was used, a leave-one-out approach was used (which tends to lead to overestimation of device accuracy), MVPA was not split into moderate and vigorous intensities, and no multi-accelerometer combinations were assessed. A follow-up study by the same group improved these limitations, using a left-wrist worn accelerometer to assess activity intensity as a four-class variable (sedentary, light, moderate, vigorous) using six different feature sets, six different machine learning models, and using a distinctly independent sample for cross validation of the models (Montoye *et al.*, 2018). Their results indicated that the random forest machine learning model coupled with a feature set comprising time-domain features (e.g., mean, standard deviation, percentiles of acceleration signal) performed optimally for determining activity intensity. Limitations to the study included no comparison of activity monitor placements or combinations.

The present study extends findings from the previous two studies in assessing activity type. Our study aim was to offer a unique comparison of the accuracy of four machine learning algorithms coupled with accelerometers worn on the wrists, hip, and ankle (along with combinations of these placements) to assess activity intensity as a two-class (MVPA or not) and as a four-class variable (sedentary, light, moderate, vigorous) using an independent-

sample cross-validation approach. The four machine learning algorithms used were decision tree, bagging, random forest, and boosting. A secondary study aim was to assess the effect of the demographic variables sex, body mass index (BMI), and age on accuracy of the developed machine learning models.

2 Methods

2.1 Participants

Thirty healthy adults (n=15 female) aged 18-79 without orthopedic limitations were recruited for this study. In order to increase variability in age and fitness level, 10 adults (n=5 female) were chosen from each of three age categories: 18-39, 40-59, and 60-79 years. Only participants had valid data 168 to be included in the study; participant demographics for those included in 169 the study can be found in Table 1. To be eligible for the study, participants had be able to perform self-paced jogging for a minimum of two minutes; in this way, all participants could theoretically perform all activities included in the study. Prior to participation, all participants had study details explained both verbally and in writing, and all gave written informed consent. Study procedures were approved by the Ball State University Institutional Review Board.

2.2 Equipment

A total of four ActiGraph GT9X Link accelerometers (ActiGraph Corp., Pensacola, FL, USA) were worn during this study. These accelerometers were worn on the dorsal aspect of the left and right wrists, over the right hip at the level of the anterior axillary line (secured using an elastic belt), and on the lateral aspect of the right ankle. All monitors were time-synchronized at the beginning of each visit and were set to record raw, triaxial accelerometer data at a sampling rate of 60 Hz. The display screen on the accelerometers was disabled; in this way, no feedback was available to participants from the monitors.

2.3 Protocol

Participants reported to the Clinical Exercise Physiology Laboratory at Ball State University for two visits, each of which took approximately two hours. Participants were instructed to arrive having not performed vigorous-intensity physical activity and not having consumed stimulants (e.g., caffeine, nicotine) or caloric food or beverages within the previous three hours. Height and weight were measured according to standardized procedures. Participants were fitted with all accelerometers as described above, and instructions were given for performance of activities for each visit.

2.3.1 Visit 1: Structured Setting

Visit 1 was designed to be highly structured, with research staff controlling many aspects of the activities. For this visit, research staff selected 11 activities (from a set of 21 possible activities) that participants were to perform within the laboratory setting. Activities were selected by the research staff in a randomized manner so that all activities within a category were performed by roughly the same number of participants. We chose these types

of activities to represent an array of activities adults might perform in a normal day. Every participant began this visit by lying supine on a padded table for 10 minutes. Afterward, we randomly assigned two additional sedentary activities, four household/chore activities, and four ambulatory/exercise activities for 5 minutes each. The activity order was meant to generally increase activity intensity throughout the visit. For the sedentary and lifestyle/chore activities, participants were instructed to perform these activities as similarly as in their everyday lives. For the ambulatory/exercise activities, participants self-selected the activity speed/intensity but were required to remain at a consistent speed/intensity for the entire activity. At the end of each activity, participants were able to take 1-2 minutes of rest before starting the next activity. Sedentary activities included reading, using a computer, watching television, writing, and playing cards; lifestyle/chore activities included standing, dusting, making a bed, folding laundry, sweeping, vacuuming, simulated gardening, and picking up items from the floor; ambulatory/exercise activities included slow and fast overground walking, self-paced treadmill walking, overground jogging, treadmill jogging, stationary cycling, and ascending/descending stairs. This structure rarely resulted in activities being performed more than once during the approximately 80 minute visit. Two researchers recorded activity start/stop and intensity of activities throughout each visit to ensure accurate criterion data were captured. Upon activity transitions, researchers discussed and came to agreement on activity intensity in real time. Then, each activity type was cross-checked in the Compendium of Physical Activities (Ainsworth *et al.*, 2011) to confirm activity intensity. This procedure was also used in Visit 2.

2.3.2 Visit 2: Semi-Structured (Simulated Free-Living) Setting

Visit 2 was designed to have less structure than Visit 1, thereby better simulating how participants might perform activities in their everyday lives. Past work has utilized such protocols in the hopes of improving generalizability of findings to free-living settings (Montoye *et al.*, 2015; Staudenmayer *et al.*, 2015). Once fitted with accelerometers, participants were asked to spend colorred approximately 80 minutes in the laboratory performing activities (from a list of the same 21 possible activities from Visit 1) as they would in their daily lives. In order to increase the variety of activities performed, participants were asked to complete at least four sedentary, four lifestyle/chore activities, and four ambulatory/exercise activities for 2-15 minutes each. Additionally, since previous research reports that adults spend the majority of their waking hours in sedentary pursuits, participants were asked to spend at least 40 minutes performing the activities in the sedentary category (Donaldson *et al.*, 2016; Matthews *et al.*, 2008). In Visit 1 we controlled the activities performed, their exact duration and order but in Visit 2 we put some general constraints within which participants were allowed to choose how much time they spent in each activity, which activities they wanted to perform, and the order of activities. Participants were also allowed to perform the same activities multiple times but this rarely resulted in activities being performed more than twice and when they were counted as only one activity. Research staff directly observed and recorded the exact start and end times of each activity during Visit 2.

2.4 Data Cleaning

Data from Visit 1 and Visit 2 were processed and cleaned in the same way and are described together. During data collection, initialization/download issues of monitors (n=1) and in-

correct orientation of monitors (n=1) were discovered in one or both visits, resulting in the exclusion of two participants' data of the 30 recruited for the study. Thus, data from 28 participants were available for model development and testing and demographic information concerning these 28 participants can be found in Table 1.

	Total Sample (n=28)	Males (n=14)	Females (n=14)
Age (years)	48.0 ± 19.6	48.5 ± 19.8	47.6 ± 20.2
Height (cm)	174.0 ± 9.0	180.4 ± 6.7	167.6 ± 5.8
Weight (kg)	80.1 ± 15.8	88.6 ± 12.5	71.5 ± 14.2
Body mass index (BMI: kg/m ^ 2)	26.3 ± 4.3	27.1 ± 3.2	25.4 ± 5.2

(a) Demographic information by sex (given as mean $\pm$ std. deviation)

Age category |	Young (18-39)	Middle (40-59)	Old (60-79)
Participants	10	9	9
BMI category |	Normal (< 25)	Overweight ($25.0 - 29.9$)	Obese (≥ 30)
Participants	11	12	5

(b) Distribution of ages and BMIs

Table 1: Participant Demographic Information

Acceleration signals from the four ActiGraph GT9X Link accelerometers were divided into nonoverlapping 30 second intervals. For each interval, the following time-domain features were computed for each axis of the accelerometer: mean, variance, minimum, maximum, upper percentiles (70th, 80th, 90th), and pairwise covariance of acceleration signals from three axes. In addition to these time-domain features, participants' age, sex, height and weight are also considered as features in the classification process.

MET values for each activity were estimated using the 2011 Compendium of Physical Activities (Ainsworth *et al.*, 2011). Then, the intensity of each activity was determined as one of four levels using the standard absolute MET thresholds of $\leq$ 1.5 METs as sedentary, 1.6-2.9 METs as light, 3.0-5.9 METs as moderate, and $\geq$ 6.0 METs as vigorous intensities (Tremblay *et al.*, 2017). We also did a sub-analysis where intensity categories were collapsed into $<$ 3.0 METs as low and $\geq$ 3.0 METs at MVPA. These directly observed activity intensities served as the criterion/ground truth for development of prediction models using accelerometer data. Once criterion data were coded according to activity intensity, they were reintegrated to 30-second epochs/windows as done in past work (Montoye *et al.*, 2015).

At the end of each visit, raw accelerometer data from each monitor were downloaded and stored as comma separated version (.csv) files. From these, features of the raw acceleration signal were extracted in 30-second epochs. Features were chosen in accordance with past work showing that the feature set provides optimal accuracy for predicting activity intensity (Montoye *et al.*, 2017, 2018). Data from all accelerometers were time-aligned with the criterion data, and all times coded as transitions between activities or breaks taken between activities in the criterion data were removed from the dataset.

In both the training and test data sets we had $n = 4313$ observations. For each of the $n = 28$ subjects we had approximately $4313/28 \approx 154$ observations. Each observation was a summary of the 30 second epoch of accelerometer data with each subject observed for approximately $154 * 30 = 4620$ seconds $= 77$ minutes. The 4313 observations from our

$n = 28$ participants provide a rich amount of variability in activity intensities with a varied population in terms of gender, age and BMI performing a range of activities as described in sections 2.3.1 and 2.3.2.

2.5 Predictive Model Development and Testing

Our study sought to answer several primary questions: 1) what single accelerometer placement has the lowest error rate for assessing activity intensity, 2) is there a combination of accelerometers that provides lower error than any single accelerometer, 3) which type of predictive model has the lowest error rate for assessing activity intensity, and 4) did accuracy of the developed machine learning models differ across demographic variables including age, sex, and/or weight status? To answer these questions, we used an independent-samples procedure. Data from Visit 1 were used as "training" data in order to develop predictive models which would classify activity intensity based on the accelerometer data. Visit 2 was used as "testing" data, where error rates of the predictive models developed from Visit 1 data would be evaluated by comparing predicted activity intensity in each 30-second interval to the criterion measure of activity intensity. In this way, there was no overlap between training and testing data. Additionally, because Visit 2 is more similar to a participants' natural behaviors than Visit 1, the performance of the models in Visit 2 is meant to gain some understanding of expected error if these models were used to assess activity intensity in a free-living setting.

We developed prediction models for four accelerometer placements (left wrist, right wrist, right hip, right ankle) as well as all possible two- and three-accelerometer combinations. Additionally, for each accelerometer/combination, four machine/statistical learning modeling methods were used: a base learner classification tree, and three ensemble learning methods including bagging, random forest, and boosting. Finally, all predictive models were developed first to categorize intensity as a four-class variable (sedentary, light, moderate, vigorous) and second as a two-class variable (low, MVPA).

Decision tree based learning methods are powerful classifiers that utilize a non-parametric tree structure to model the relationship between a feature set and the outcome. Classification tree algorithms may be divided into two groups: 1) trees based on binary recursive partitioning and 2) trees based on nonbinary or multiway splits (Kim and Loh, 2001). The common theme of these algorithms is to split the feature space into subsets, which are then split repeatedly into smaller subsets, until the process stops when some node impurity conditions are satisfied. Four recursive binary tree algorithms such as a base classification tree, bagging, random forest, and boosting are implemented to address questions 1 and 2 above. Then classification accuracies of these four algorithms are compared for the best performing device placement in terms of lowest misclassification error. Finally, the best algorithm with the best placements are assessed to explore prediction accuracies among subgroups of participants in terms of age, sex and weight status.

2.5.1 Single Classification Trees

Classification trees are the part of a larger group of decision tree-based models known as classification and regression trees (CART) (Breiman *et al.*, 1984) used for making predictions about outcomes of interest. CART models are built recursively with binary partitions of feature spaces. An optimal classification tree is built upon utilizing certain optimiza-

tion criteria, such as maximizing accuracy or minimizing misclassification. Another related algorithm for building a classification tree is the C5.0 (Quinlan, 1993) algorithm. Both algorithms generate very similar classification results although the optimization criteria differs during the splitting of the feature space. A single classification tree based on CART or the C5.0 principle serves as a good reference method for more complex ensemble tree-based methods for classification and is considered as a base learner in the process of systematic selection of accelerometer placement.

2.5.2 Ensemble Methods

We compared the performance of three ensemble decision tree classification methods against the reference single classification tree. These are bootstrap aggregating or bagging (Breiman, 1996), random forest (Breiman, 2001), and boosting (Freund and Schapire, 1997). Both bagging and random forest algorithms are based on single trees built on bootstrap training samples. However, unlike bagging, in random forest, each tree is constructed by taking only a random sample of predictors without replacement before each node is split. The main idea behind the third ensemble algorithm is to sequentially apply a weak classifier (whose error rate is only slightly better than random guessing) to repeatedly modified versions of the data. Predictions from this sequence of weak classifiers are then combined through a weighted majority vote to produce a final prediction. While the random forest algorithm has been widely applied in many classification applications, implementation of the bagging and boosting algorithms is somewhat limited in particular for classifying physical activity intensity levels. In the current study, we implement all four decision tree-based classifiers to address our primary research questions as well as to assess relative performance of these algorithms for the best accelerometer placement in terms of error rates.

The four classification algorithms, single tree, bagging, random forest, and boosting are trained on Visit 1 data using summary statistics of frequency measures from Visit 1 participants as our training feature set and physical activity intensity levels in Visit 1. Prediction accuracies of all the classification methods are computed treating Visit 2 data as the independent test sample.

2.5.3 Prediction Error Rate Analysis

To make comparisons of error rates between different combinations of locations and between different classification methods in sections 3.1 and 3.2, we make comparisons with a reference combination of locations and a reference classification method, respectively. 1200 bootstrap samples were created by sampling 1200 times with replacement from all $n = 4313$ observations in the training set, building models on each bootstrap sample and computing classification errors for each of the 1200 samples on the testing set. Differences between classification errors were computed for each individual sample to remove effects introduced by different subjects in the study. Medians of error rates for the reference and medians of differences of error rates with the reference over the 1200 samples were computed. Using the percentile method, approximate 95% bootstrap confidence intervals (CIs) were constructed as in Hogg *et al.* (2005), with a Bonferroni correction to account for multiple comparisons. Point estimates are given as classification errors from the model built on the training set or as medians of classification errors from the bootstrap samples. In sections 3.1.1 and 3.2.1,

the boosting algorithm is applied to training sets for different combinations of locations and confusion matrices are built on test sets.

We stratified the results on demographic variables to look for effects on accuracy. We also computed kappa values to compare observed versus expected accuracies and to account for correct classification by chance. For classification of four activity levels we also computed weighted kappa values (Cohen, 1968) as the data is ordinal and misclassifications further away from true classifications are more significant than misclassifications closer to true classifications. For example, we want to penalize a misclassification of sedentary as light as less significant than a misclassification of sedentary as moderate or vigorous. We use a linear weighted kappa measure to do so, with movement from the true classification from one class to another penalized equally. A scale formulated in Landis and Koch (1977) and often used to interpret kappa values is given in Table 2.

Kappa	Agreement
0.01-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Good
0.61-0.80	Substantial
0.81-0.99	Almost perfect

Table 2: Interpreting Kappa

3 Results

For each predictive model, a test classification error from the test data on Visit 2, a bootstrap median test error, and a 95% Bonferroni corrected bootstrap confidence interval were reported for four accelerometer placements (left wrist, right wrist, right hip, right ankle) as well as all possible single, two- and three-accelerometer combinations. Our first objective was to determine: 1) a single accelerometer placement with lowest classification error for assessing activity intensity, and 2) a combination of accelerometer placement with the lowest error rate for assessing activity intensity. Test statistics are generated and these goals are explored for both four-class (sedentary, light, moderate, vigorous) prediction in section 3.1 and for two-class (low, MVPA) prediction in section 3.2. For the best single and combined accelerometer placements (in terms of error rates), we identified the predictive models with lowest error rates, again considering both four- and two-class prediction. Finally, we explored if the accuracy of activity intensity prediction varied across a selected number of demographic variables (age, sex, and BMI) for the best predictive model. Data from four study participants and all the code used to create the models, tables and figures in this paper is available at https://github.com/DrewLazar/SL_Accelerometers.

3.1 Four-Class Prediction

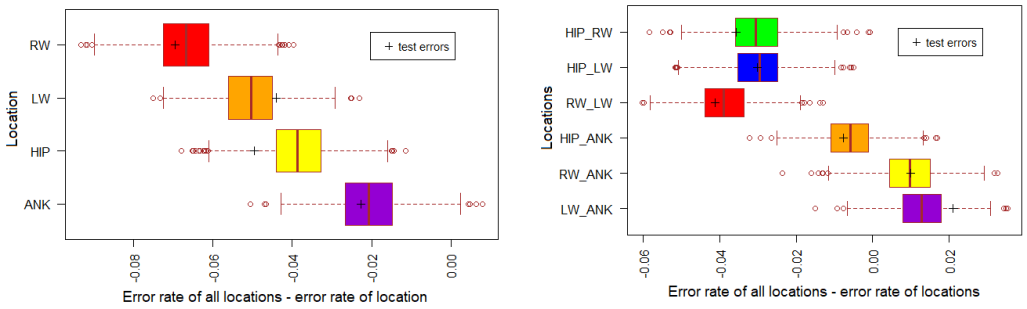
In assessing error for four-class prediction (sedentary, light, moderate, vigorous) for bagging, any single- or multi-accelerometer placement which included the ankle-worn accelerometer had errors not significantly different from the reference placement (combination of all four accelerometers), whereas all other single- or multi-accelerometers placements which did not include the ankle had significantly higher error rates than the reference placement (Table 3a).

Similarly, for boosting and random forest, all two- and three-accelerometer combinations including the ankle were not different from the reference placement (with the exception of the ankle-left wrist accelerometer combination, where error was significantly lower than the

reference), whereas all other two- and three-accelerometer combinations which did not include the ankle had significantly higher error than the reference placement (Tables 3b, 3c). However, all single-accelerometer placements had significantly higher error than the reference placement for boosting and random forest models (Tables 3b, 3c).

For the single classification tree, the right wrist accelerometer had significantly higher error than the reference method, but none of the other single accelerometer placements were significantly different from the reference method. In addition, errors for all two- and three-accelerometer combinations were not significantly different from the reference method (Table 3d).

Figure 1 presents, for the boosting algorithm, distributions of differences of errors of single- and two-accelerometer placements with the multi-accelerometer reference placement. In comparison with single- vs. multi-accelerometer prediction, the two-accelerometer ankle-left wrist combination and ankle-right wrist combination had slightly but consistently lower error point estimates than any single-accelerometer placement for bagging, boosting, and random forest with error point estimates 2.85-4.38 percentage points lower compared to the ankle accelerometer with these three modeling methods (Tables 3a, 3b 3c). Conversely, for the classification tree, the ankle accelerometer placement had an error 1.09-1.40 percentage points lower than the ankle-left wrist and ankle-right wrist combinations (Table 3d).



(a) One Placement (b) Two Placements

ANK=ankle, HIP=hip, LW=left wrist, RW=right wrist

Figure 1: Distributions of differences of prediction errors between one and two placements and all placements for boosting algorithm - four activity levels.

Since the ankle accelerometer placement had the lowest predictive error of the single accelerometers, and since multi-accelerometer combinations using the ankle (i.e., ankle-left wrist and ankle-hip) had the lowest overall error, these were used to compare accelerometer modeling methods using the classification tree as the reference method (Table 4). Tree classification is the reference method as it is a base learner as explained in section 2.5. For classification using the ankle accelerometer or the ankle-hip combination, error was not different among modeling methods. However, for the ankle-left wrist combination, the classification tree had significantly higher error than all other classification methods. For the ankle accelerometer and two-accelerometer combinations, point estimates were slightly lower for boosting compared to the other classification methods (Figure 2).

Error rate of all locations - error rate of location(s). Error rates: Lower Same Higher

# of Locs	Locs	Test Error	Median	95% C.I.
4	all	0.305	0.312	(0.297, 0.331)
1	ak	-0.027	-0.021	(-0.045, 0.008)
	hp	-0.049	-0.047	(-0.074,-0.019)
	lw	-0.045	-0.048	(-0.074,-0.021)
	rw	-0.060	-0.067	(-0.095,-0.039)
2	lw&ak	0.016	0.010	(-0.014, 0.038)
	rw&ak	0.008	0.008	(-0.016, 0.033)
	hp&ak	-0.017	-0.014	(-0.039, 0.010)
	rw&lw	-0.042	-0.045	(-0.070,-0.018)
	hp&lw	-0.033	-0.033	(-0.053,-0.011)
	hp&rw	-0.040	-0.034	(-0.057,-0.012)
3	hp&lw&rw	-0.036	-0.031	(-0.047,-0.014)
	ak&lw&rw	0.003	0.005	(-0.015, 0.031)
	an&hp&rw	0.002	0.001	(-0.018, 0.018)
	hp&ak&lw	-0.001	-0.001	(-0.015, 0.015)

(a) Bagging

# of Locs	Locs	Test Error	Median	95% C.I.
4	all	0.304	0.312	(0.300, 0.325)
1	ak	-0.023	-0.021	(-0.040,-0.001)
	hp	-0.049	-0.039	(-0.062,-0.017)
	lw	-0.044	-0.050	(-0.071,-0.030)
	rw	-0.070	-0.067	(-0.089,-0.044)
2	lw&ak	0.021	0.013	(-0.006, 0.030)
	rw&ak	0.010	0.009	(-0.012, 0.028)
	hip&ak	-0.008	-0.006	(-0.025, 0.012)
	rw&lw	-0.041	-0.039	(-0.058,-0.019)
	hip&lw	-0.030	-0.029	(-0.050,-0.010)
	hip&rw	-0.036	-0.031	(-0.050,-0.009)
3	hp&lw&rw	-0.028	-0.029	(-0.047,-0.014)
	ak&lw&rw	0.011	0.008	(-0.015, 0.031)
	ak&hp&rw	0.004	0.002	(-0.018, 0.018)
	hp&ak&lw	0.010	0.000	(-0.015, 0.014)

(b) Boosting

# of Locs	Locs	Test Error	Median	95% C.I.
4	all	0.314	0.315	(0.306, 0.323)
1	ank	-0.020	-0.024	(-0.049,-0.009)
	hp	-0.052	-0.048	(-0.065,-0.034)
	lw	-0.037	-0.042	(-0.056,-0.027)
	rw	-0.059	-0.065	(-0.082,-0.050)
2	lw&ak	0.014	0.013	(0.000, 0.026)
	rw&ak	0.009	0.006	(-0.007, 0.018)
	hp&ak	-0.010	-0.008	(-0.023, 0.003)
	rw&lw	-0.042	-0.040	(-0.055,-0.027)
	hp&lw	-0.034	-0.030	(-0.043,-0.017)
	hp&rw	-0.035	-0.036	(-0.050,-0.024)
3	hp&lw&rw	-0.033	-0.030	(-0.040,-0.019)
	ak&lw&rw	0.003	0.007	(-0.005, 0.017)
	ak&hp&rw	0.007	0.002	(-0.009, 0.001)
	hp&ak&lw	0.004	0.001	(-0.001, 0.010)

(c) Random Forest

# of Locs	Locs	Test Error	Median	95% C.I.
4	all	0.333	0.336	(0.308, 0.365)
1	ak	-0.019	-0.021	(-0.121, 0.030)
	hp	-0.041	-0.040	(-0.096, 0.009)
	lw	-0.031	-0.037	(-0.089, 0.003)
	rw	-0.064	-0.082	(-0.173,-0.016)
2	lw&ak	-0.033	-0.008	(-0.054, 0.042)
	rw&ak	-0.008	0.001	(-0.036, 0.047)
	hip&ak	-0.025	-0.014	(-0.050, 0.029)
	rw&lw	-0.021	-0.032	(-0.106, 0.011)
	hp&lw	-0.006	-0.028	(-0.010, 0.011)
	hp&rw	-0.016	-0.021	(-0.077, 0.028)
3	hp&lw&rw	-0.004	-0.020	(-0.075, 0.013)
	ak&lw&rw	-0.008	0.001	(-0.036, 0.044)
	ak&hp&rw	0.000	0.000	(-0.026, 0.031)
	hp&ak&lw	-0.003	0.000	(-0.043, 0.019)

(d) Tree

Locs=Locations, ak=ankle, hp=hip, lw=left wrist, rw=right wrist

Table 3: Comparison of prediction accuracies of activity intensity classifications of different placements versus all placements - four activity levels

3.1.1 Comparisons Across Demographic Variables for Boosting Method for Four Classes

Finally, comparisons across the demographic variables of age, weight status, and sex were made, using the ankle accelerometer and two-accelerometer ankle-left wrist and ankle-hip combinations with the boosting classification method as these provided the lowest classification errors in the previous analyses. Confusion matrices were constructed for each demographic variable separately considering accelerometer placements on ankle, ankle-left wrist combination, and ankle-hip combination using prediction results from the boosting classification method. Table 5 presents classification errors and two agreement measures, kappa and weighted kappa as given in 2.5.3, based on these confusion matrices.

When stratified by age (Table 5a), classification errors decreased, and kappa and weighted kappa scores generally increased, with increasing age. In examining the confusion matrices, the improved accuracy in older individuals may be partly due to higher participation in sedentary behaviors and lower participation in vigorous activities in the oldest

Error rate of tree method - error rate of method
Error rates: Lower Same Higher

Locs	Method	Test Error	Median	95% C.I.
Ank	tree	0.351	0.3520	(0.325, 0.442)
	bag	0.019	0.022	(-0.017, 0.113)
	boost	0.026	0.025	(-0.017, 0.120)
	rf	0.016	0.017	(-0.020, 0.110)
Ank&lw	tree	0.366	0.343	(0.318, 0.372)
	bag	0.074	0.042	(0.014, 0.072)
	boost	0.083	0.044	(0.017, 0.074)
	rf	0.071	0.042	(0.016, 0.073)
Ank&hip	tree	0.357	0.352	(0.314, 0.384)
	bag	0.034	0.025	(-0.017, 0.062)
	boost	0.046	0.033	(-0.009, 0.065)
	rf	0.036	0.028	(-0.007, 0.061)

tree=decision tree, bag=bagging, boost=boosting, rf=random forest, Ank=ankle, lw=left wrist

Table 4: Comparison of classification methods to tree method with ankle, ankle & left wrist, and ankle & hip placements - four activity levels

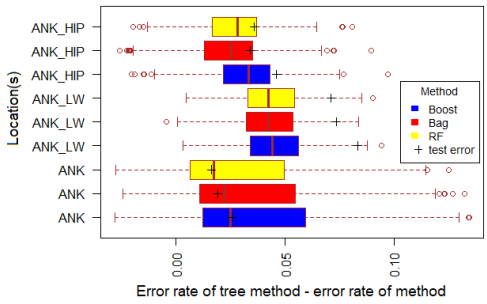


Figure 2: Distributions of differences of classification errors between different methods and tree method - four activity levels

group compared with the other two groups. For middle- and older-aged adults, weighted kappa scores indicated substantial agreement, and for younger-aged adults, weighted kappa scores indicated good agreement for the ankle accelerometer and both two-accelerometer combinations.

When stratified by weight status (Table 5b), classification errors were consistently highest in the normal-weight group and lowest in the obese group. Kappa and weighted kappa scores were not consistently different across groups, although the obese group had the highest kappa and weighted kappa scores for both two-accelerometer combinations and overweight had the highest kappa and weighted kappa scores for the ankle accelerometer. The confusion matrices specific to weight status revealed virtually no participation in vigorous-intensity physical activity in the obese group and may partially explain the higher accuracy for intensity prediction in this group. Weighted kappa scores indicated substantial agreement in the overweight and obese groups for the ankle-left wrist combination and good agreement for all groups with the other accelerometers/combinations.

Finally, when stratified by sex (Table 5c), classification errors were lower and kappa and weighted kappa scores were higher for activity intensity prediction in males compared to females for the ankle accelerometer and both two-accelerometer combinations. For both females and males, weighted kappa scores indicated substantial agreement for the ankle accelerometer and both two-accelerometer combinations.

Figure 3 presents agreement plots as given in Bangdiwala (2017). These plots are for the accelerometer placements (ankle and left wrist) and demographic strata with the lowest classification errors and highest kappa values. For each activity level, the width and height of the outer rectangle gives the marginal number of classifications of the activity level by the boosting algorithm and true classification, respectively. The width of the black inner square is the number of agreements for a particular class. The difference of the heights of the outer rectangle to the black inner square is the number of misclassifications of that particular class and the difference in the widths is the number of misclassifications of other classes as that particular class. The gray shading represents misclassifications of adjacent

Locs	Variable	Error Rate	Kappa	Weighted Kappa
Ank	Overall	0.3263	0.5255	0.6578
	Young	0.3824	0.4527	0.5559
	Middle	0.3010	0.5596	0.7044
	Old	0.2965	0.5599	0.7049
Ank &lw	Overall	0.2824	0.5860	0.7046
	Young	0.3641	0.4820	0.5801
	Middle	0.2686	0.6035	0.7414
	Old	0.2211	0.6612	0.7788
Ank &hip	Overall	0.3112	0.5401	0.6725
	Young	0.3950	0.4366	0.5492
	Middle	0.2860	0.5730	0.7196
	Old	0.2569	0.6068	0.7399

(a) Stratified by Age: Young (18-39), Middle (40-59), Old (60-79)

Locs	Variable	Error Rate	Kappa	Weighted Kappa
Ank	Overall	0.3263	0.5255	0.6578
	Normal	0.3574	0.4851	0.6275
	Overweight	0.3113	0.5500	0.6764
	Obese	0.2900	0.5372	0.6531
Ank &lw	Overall	0.2824	0.5860	0.7046
	Normal	0.3080	0.5562	0.6859
	Overweight	0.2876	0.5811	0.6764
	Obese	0.2133	0.6526	0.7498
Ank &hip	Overall	0.3112	0.5401	0.6725
	Normal	0.3313	0.5212	0.6597
	Overweight	0.3141	0.5380	0.6719
	Obese	0.2593	0.5759	0.6873

(b) Stratified by BMI (kg/m^2): Normal (< 25.0), Overweight ($25.0 - 29.9$), Obese (≥ 30)

Locatic	Variable	Error Rate	Kappa	Weighted Kappa
Ank	Overall	0.3263	0.5255	0.6578
	Male	0.2973	0.4924	0.7023
	Female	0.3528	0.5612	0.6160
Ank &lw	Overall	0.2824	0.5860	0.7046
	Male	0.2706	0.5754	0.7314
	Female	0.2933	0.5967	0.6796
Ank &hip	Overall	0.3112	0.5401	0.6725
	Male	0.2987	0.5285	0.6996
	Female	0.3226	0.5527	0.6478

(c) Stratified by Sex

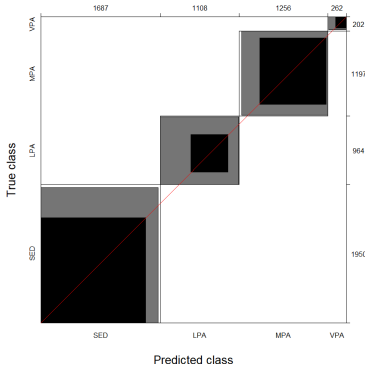
Table 5: Predictive accuracy of boosting method for ankle, ankle & left wrist, and ankle & hip placements stratified by demographic variables - four classes

classes. We can see in these plots that misclassifications outside of adjacent classes for all demographic categories are relatively few. In the obese class there are only four participants who engaged in vigorous activity, one of which was correctly classified. In general, the classes aren't balanced with vigorous activity the least common, then light activity, followed by moderate activity and then sedentary activity the most common in line with the structure of Visit 2 as given in section 2.3.2.

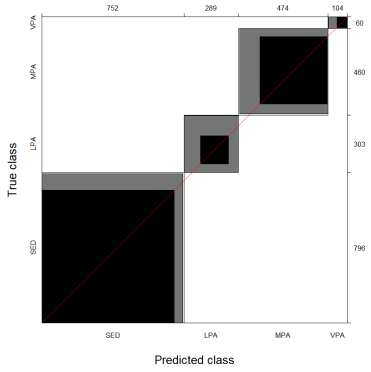
3.2 Two-Class Prediction

In assessing error for two-class prediction (MVPA vs. low), similar results were observed as with the four-class prediction. Specifically, for bagging and boosting, any single-accelerometer placement or multi-accelerometer combination which included the ankle-worn accelerometer had errors not significantly different from the reference placement of all locations, whereas all other single-accelerometer placements or multi-accelerometer combinations which did not include the ankle had significantly higher error rates than the reference placement (Tables 6a, 6b).

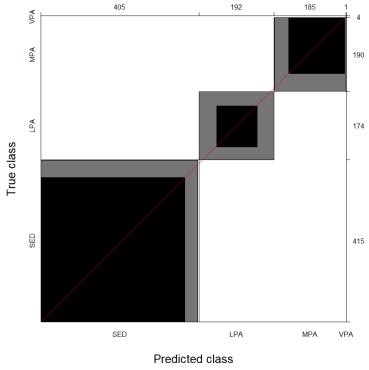
Plots for boosting indicate little difference in single-accelerometer placements for the ankle, hip, and left wrist placements but illustrate a trend for higher accuracy in two-accelerometer combinations which include the ankle (Figures 4a and 4b).



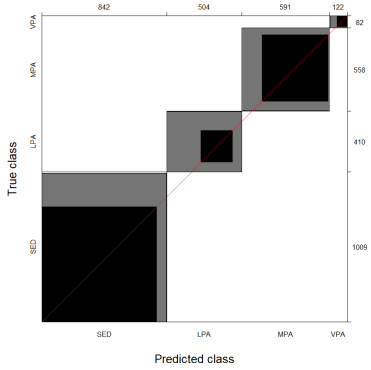
(a) Agreement Plot for All



(b) Agreement Plot for Old Age

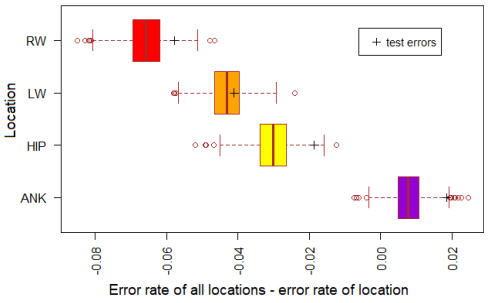


(c) Agreement Plot for Obese (BMI $\geq$ 30)

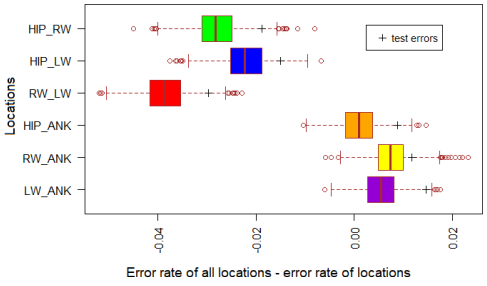


(d) Agreement Plot for Male

Figure 3: Agreement Plots for Boosting: Ankle and Left Wrist Placements



(a) One Location



(b) Two Locations

Figure 4: Distributions of differences of prediction errors between one and two placements and all placements for boosting algorithm - two activity levels

For random forest, the single-accelerometer placement on the ankle had significantly lower error than the reference placement, and all two- and three-accelerometer combinations had errors not significantly different from the reference. All single accelerometers

Error rate of all locations - error rate of location(s). Error rates: Lower Same Higher

# of Locs	Locs	Test Error	Median	95% C.I.
4	all	0.122	0.126	(0.117, 0.136)
1	ak	0.001	0.005	(-0.007, 0.018)
	hp	-0.035	-0.035	(-0.057, -0.018)
	lw	-0.045	-0.042	(-0.061, -0.025)
	rw	-0.067	-0.071	(-0.088, -0.053)
2	lw&ak	0.003	0.004	(-0.007, 0.016)
	rw&ak	0.003	0.002	(-0.008, 0.016)
	hip&ak	-0.007	-0.002	(-0.012, 0.010)
	rw&lw	-0.039	-0.038	(-0.056, -0.019)
	hp&lw	-0.019	-0.025	(-0.044, -0.007)
	hp&rw	-0.034	-0.033	(-0.051, -0.017)
3	hp&lw&rw	-0.024	-0.025	(-0.042, -0.008)
	ak&lw&rw	0.001	0.002	(-0.008, 0.013)
	ak&hp&rw	0.001	-0.001	(-0.008, 0.007)
	hp&ak&lw	-0.002	0.001	(-0.007, 0.009)

(a) Bagging

# of Locs	Locs	Test Error	Median	95% C.I.
4	all	0.129	0.124	(0.118, 0.132)
1	ank	0.019	0.008	(-0.003, 0.019)
	hp	-0.019	-0.030	(-0.044, -0.019)
	lw	-0.041	-0.043	(-0.055, -0.031)
	rw	-0.058	-0.066	(-0.080, -0.053)
2	lw&ak	0.015	0.005	(-0.004, 0.016)
	rw&ak	0.012	0.007	(-0.002, 0.019)
	hip&ak	0.009	0.001	(-0.009, 0.012)
	rw&lw	-0.030	-0.039	(-0.050, -0.026)
	hp&lw	-0.015	-0.022	(-0.035, -0.011)
	hp&rw	-0.019	-0.028	(-0.040, -0.015)
3	hp&lw&rw	-0.017	-0.024	(-0.042, -0.008)
	ak&lw&rw	0.011	0.004	(-0.008, 0.013)
	ak&hp&rw	0.004	0.001	(-0.008, 0.007)
	hp&ak&lw	0.001	0.001	(-0.007, 0.009)

(b) Boosting

# of Locs	Locs	Test Error	Median	95% C.I.
4	all	0.126	0.125	(0.120, 0.131)
1	ak	0.007	0.009	(0.000, 0.017)
	hp	-0.031	-0.032	(-0.043, -0.021)
	lw	-0.038	-0.040	(-0.050, -0.029)
	rw	-0.066	-0.065	(-0.076, -0.055)
2	lw&ak	0.006	0.007	(-0.001, 0.014)
	rw&ak	0.008	0.006	(-0.001, 0.013)
	hip&ak	0.001	0.002	(-0.007, 0.008)
	rw&lw	-0.036	-0.038	(-0.048, -0.027)
	hp&lw	-0.026	-0.024	(-0.032, -0.015)
	hp&rw	-0.030	-0.031	(-0.040, -0.022)
3	hp&lw&rw	-0.031	-0.026	(-0.035, -0.018)
	ak&lw&rw	0.005	0.005	(-0.001, 0.012)
	ak&hp&rw	0.001	0.000	(-0.005, 0.005)
	hp&ak&lw	-0.001	0.001	(-0.005, 0.006)

(c) Random Forest

# of Locs	Locs	Test Error	Median	95% C.I.
4	all	0.133	0.134	(0.117, 0.152)
1	ak	0.000	0.000	(-0.026, 0.021)
	hp	-0.040	-0.030	(-0.076, 0.005)
	lw	-0.050	-0.042	(-0.081, -0.015)
	rw	-0.057	-0.073	(-0.119, -0.036)
2	lw&ak	0.000	0.000	(-0.024, 0.025)
	rw&ak	0.000	0.000	(-0.024, 0.025)
	hp&ak	0.000	0.000	(-0.025, 0.012)
	rw&lw	-0.044	-0.037	(-0.079, -0.007)
	hp&lw	-0.031	-0.025	(-0.074, 0.009)
	hp&rw	-0.036	-0.029	(-0.068, 0.008)
3	hp&lw&rw	-0.024	-0.027	(-0.067, 0.006)
	ak&lw&rw	0.000	0.000	(-0.026, 0.018)
	ak&hp&rw	0.000	0.000	(-0.007, 0.004)
	hp&ak&lw	0.000	0.000	(-0.012, 0.007)

(d) Tree

Locs=Locations, ak=ankle, hp=hip, lw=left wrist, rw=right wrist

Table 6: Comparison of prediction accuracies of activity intensity classifications of different placements versus all placements - two activity levels

other than the ankle and multi-accelerometer combinations not including the ankle had significantly higher error than the reference placement (Table 6c).

Finally, for the classification tree modelling approach, the single-accelerometer placements on the ankle and hip, all two-accelerometer combinations which included the ankle, and all three-accelerometer combinations were not different from the reference, whereas the single-accelerometer placement on either wrist as well as the two-accelerometer placement with both wrists had significantly higher error than the reference placement (Table 6d).

In comparing modeling methods with the best-performing accelerometer placements including only on the ankle and two-accelerometer combinations of ankle-left wrist and ankle-hip (Table 7), boosting and random forest had significantly lower errors than classification tree for the ankle accelerometer and ankle-left wrist combination. There were no other significant differences among model types, although point estimates for boosting indicated slightly lower errors than the other modeling methods for the ankle and both two-accelerometer combinations tested (Figure 5).

Error rate of tree method - error rate of method
Error rates: Lower Same Higher

Locs	Method	Test Error	Mediar	95% C.I.
Ank	tree	0.133	0.135	(0.125, 0.168)
	bag	0.015	0.014	(-0.000, 0.049)
	boost	0.023	0.018	(0.004, 0.051)
	rf	0.017	0.018	(0.007, 0.052)
Ank&lw	tree	0.133	0.135	(0.121, 0.161)
	bag	0.012	0.013	(-0.003, 0.044)
	boost	0.019	0.016	(0.002, 0.047)
	rf	0.014	0.016	(0.003, 0.043)
Ank&hip	tree	0.133	0.135	(0.115, 0.162)
	bag	0.001	0.006	(-0.016, 0.036)
	boost	0.013	0.011	(-0.012, 0.038)
	rf	0.010	0.011	(-0.009, 0.038)

tree=decision tree, bag=bagging, boost=boosting, rf=random forest, Ank=ankle, lw=left wrist

Table 7: Comparison of classification methods to tree method with ankle, ankle & left wrist, and ankle & hip placements - two activity levels

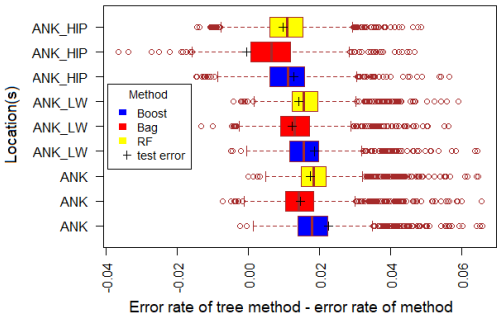


Figure 5: Distributions of differences of classification errors between different methods and tree method - two activity levels

3.2.1 Comparisons Across Demographic Variables for Boosting Method for Two Classes

Finally, subanalyses by age, weight status, and sex were performed as shown in Table 8. When stratifying by age, the classification error was highest (and kappa lowest) in the youngest age group, and classification error was lowest (and kappa highest) in the middle age group. Kappa scores indicated good agreement adjusted for agreement by chance alone for all age groups and all accelerometers/combinations.

Stratifying by weight status, classification errors were lowest in the obese group for the ankle and ankle-left wrist combination, although errors were lowest in the normal weight group with the ankle-hip combination. Conversely, kappa scores were highest for the normal weight group for all accelerometers/combinations. The discord between the kappa and classification error scores is likely due to the proportion of time spent in each intensity category, with the obese group spending substantially less time in MVPA (23.8%) than the youngest group (35.5%) and overweight group (39.9%). As with the other stratifications, kappa scores indicated good agreement across all weight status groups.

Similar to the four-class analysis, males had lower error and higher kappa scores than females for the ankle accelerometer and both two-accelerometer combinations, and kappa scores for both sexes indicated good agreement.

4 Discussion

Our study sought to answer several interrelated questions, including determination of the most accurate single- and multi-accelerometer placements and the highest performing predictive model for assessing physical activity intensity in an adult population diverse in age and fitness level. Second, we conducted subgroup analyses to determine if accuracy of the developed models varied according to age, sex, or weight status. This subanalysis, and in

Locs	Variable	Error Rate	Kappa
Ank	Overall	0.1104	0.7523
	Young	0.1338	0.6982
	Middle	0.0964	0.7841
	Old	0.1007	0.7749
Ank&lw	Overall	0.1143	0.7425
	Young	0.1317	0.7008
	Middle	0.0940	0.7902
	Old	0.1149	0.7414
Ank&hip	Overall	0.1201	0.7302
	Young	0.1387	0.6852
	Middle	0.1043	0.7669
	Old	0.1161	0.7406

(a) Stratified by Age: Young (18-39), Middle (40-59), Old (60-79)

Locs	Variable	Error Rate	Kappa
Ank	Overall	0.1104	0.7523
	Normal	0.1097	0.7563
	Overweight	0.1187	0.7477
	Obese	0.0933	0.7322
Ank&lw	Overall	0.1143	0.7425
	Normal	0.1108	0.7508
	Overweight	0.1215	0.7407
	Obese	0.1061	0.7079
Ank&hip	Overall	0.1201	0.7302
	Normal	0.1097	0.7552
	Overweight	0.1308	0.7151
	Obese	0.1149	0.6951

(b) Stratified by BMI (kg/m^2): Normal (< 25.0), Overweight ($25.0 - 29.9$), Obese (≥ 30)

Locs	Variable	Error Rate	Kappa
Ank	Overall	0.1104	0.7523
	Male	0.0937	0.7870
	Female	0.1256	0.7213
Ank&lw	Overall	0.1143	0.7425
	Male	0.0986	0.7750
	Female	0.1287	0.7133
Ank&hip	Overall	0.1201	0.7302
	Male	0.1059	0.7591
	Female	0.1331	0.7043

(c) Stratified by Sex

Table 8: Predictive accuracy of boosting method for ankle, ankle & left wrist, and ankle & hip placements stratified by demographic variables - two classes

particular, the comparisons of placement accuracies is a major difference with the study in Montoye *et al.* (2018) and an important and novel contribution to the literature in this field in general.

Overall, our results suggest that a single accelerometer worn on the ankle is superior to hip- or wrist-worn accelerometers. Past research using machine learning methods to predict activity intensity found that a thigh-worn accelerometer had higher accuracy than hip- or wrist-worn accelerometers (Montoye *et al.*, 2016), however, less comfort and lower compliance with the thigh accelerometer placement has has also been reported. Similarly, a recent study examining accuracy of different activity monitor placement sites for step-counting showed a clear improvement in accuracy using the ankle-worn StepWatch activity monitor compared to any hip- or wrist-worn devices (Toth *et al.*, 2018). Finally, researchers examining energy expenditure prediction accuracy from activity monitors found that a monitor worn on the shoe had lower error than hip- or wrist-worn devices as well as a five-accelerometer system (Dannecker *et al.*, 2013). Therefore, our study's findings are in concordance with other recent work and indicate that activity monitors worn somewhere on the lower limb provide better assessment of physical activity than devices worn on other body locations. However, accuracy of the ankle- or thigh-placement sites must be considered alongside with compliance wearing a device at one of those locations. Thigh-worn accelerometers must be taped in place, which may be less comfortable for wearers of the device. While ankle-worn devices can be secured via elastic band, there have been anecdotal reports that ankle-worn devices are mistaken for police-issued monitoring devices, although

a recent study found high compliance wearing an ankle-worn accelerometer and specifically noted that only 2 of 459 participants refused to wear the accelerometer possibly due to concerns about perceptions associated with an ankle-worn device (Hager *et al.*, 2015). Additionally, as technology improves and devices continue to shrink in size, thigh- or ankle-worn devices may be able to be embedded into clothing (such as pants or socks) or worn as a small patch like a bandage, which would reduce burden to the wearer and make the devices less conspicuous to wear.

Aside from the ankle-worn accelerometer, the left wrist-worn accelerometer had similar error to the hip-worn accelerometer, both of which had lower error estimates than the right wrist-worn accelerometer. For all but two participants, the left wrist was their non-dominant wrist, suggesting that placement on the non-dominant wrist may yield slightly higher accuracy for assessment of activity intensity using our predictive models. This finding is in concordance with the National Health and Nutrition Examination Survey (NHANES) physical activity surveillance protocol, which has been collecting accelerometer data on participants' non-dominant wrist starting in the 2011-2014 collection cycle (Troiano *et al.*, 2014). Given the popularity of wrist-worn activity tracking devices and high compliance noted by Troiano *et al.* (2014) with wrist-worn accelerometers, in some situations the loss in accuracy with a wrist-worn device may be justified given the improved compliance.

In comparing single- with multi-accelerometer predictions, slight improvements were noted using a two accelerometer ankle-left wrist combination, although the improvements in accuracy were not seen across all modeling types. Past research has noted improvements in accuracy of multi-accelerometer systems over single accelerometers for assessment of activity type (Dong *et al.*, 2013), with one recent study by Chowdhury et al. (Chowdhury *et al.*, 2017) indicating that an ankle-wrist accelerometer combination was superior to other two- and three-accelerometer combinations tested. Conversely, evidence of potential benefits of multi-accelerometer systems for accuracy in assessing energy expenditure is less consistent (Dannecker *et al.*, 2013; Löf *et al.*, 2013). Given these past studies along with the present study's findings, the small benefit of using two accelerometers vs. one may not be worth the increased burden on wearers, although as previously noted device miniaturization and embedment in clothing may make multi-accelerometer systems less burdensome. Additionally, the ankle-left wrist combination was in some cases more accurate than any three- or four-accelerometer combination, suggesting that more accelerometer data does not necessarily improve accuracy in physical activity intensity assessment. Similar findings were shown by Mackintosh *et al.* (2016), who found that two-accelerometer systems had superior accuracy to systems of 3-8 accelerometers when assessing energy expenditure in children. As an alternate to using more accelerometers, using additional sensors alongside an accelerometer (within a single monitor or as a multi-monitor system), such as heart rate or gyroscope, has been shown to improve measurement of energy expenditure and may be considered as an option for further improvement of physical activity intensity assessment (O'Driscoll *et al.*, 2018; Lu *et al.*, 2018; Hibbing *et al.*, 2018).

Another of our primary research questions was to compare modeling methods, and we did this using our best performing single-accelerometer placement (ankle) as well as our best performing two-accelerometer placements (ankle-left wrist and ankle-hip). All ensemble methods (bagging, boosting, and random forest) had significantly lower error than the classification tree for the four-class intensity classification with the ankle-left wrist accelerometer combination and nonsignificantly trended in this direction for the ankle and ankle-hip combination. For two-class intensity classification, boosting and random forest

had significantly lower errors than the classification tree for the ankle and ankle-left wrist combination, whereas errors from bagging were not different from the classification tree. The superiority of ensemble methods to single classifiers is not surprising given the theoretical basis for the methods and is supported by recent work (Chowdhury *et al.*, 2017; Montoye *et al.*, 2018). The relatively small differences in accuracy among methods allows for some flexibility in model choice depending on computation complexity/run time desired for the modeling approaches. The similarity in accuracy of the tested ensemble methods, along with past work suggesting that feature selection from accelerometer data has minimal effect on activity intensity accuracy (Montoye *et al.*, 2018) and our present findings that additional accelerometers have minimal effect on accuracy, suggests that instead of searching for more accurate modeling methods or feature sets, improvements in accuracy for assessment of physical activity intensity may require a greater variety of sensors (e.g., heart rate) in addition to accelerometry within one or more body-worn devices. This possibility should be explored in future work.

Error rates of our best performing models were approximately 29-30% for our best performing accelerometers in the four-class intensity classification. Comparisons to past work are difficult given the nature and variety of activities performed during protocols, but our errors are slightly higher compared to the $\approx 22 - 23\%$ errors found for four-class intensity prediction in a previous study by Montoye et al. (Montoye *et al.*, 2018) using a left-wrist accelerometer and random forest modeling approach. Conversely, our errors are similar to or slightly lower than those of Sasaki et al. (Sasaki *et al.*, 2016) who found error rates of 33% for an ankle-worn accelerometer and 39% for a wrist-worn accelerometer when assessing five-class activity prediction using a random forest classifier. While Sasaki *et al.* (2016) note that the more recent machine learning approaches have helped to reduce classification error using accelerometers, the high errors for physical activity prediction necessitates further research that examines approaches to continue to improve activity intensity assessment since activity intensity is a central component of many national physical activity recommendations (Piercy *et al.*, 2018).

Subgroup analyses indicate that the models performed with substantial agreement across all age, sex, and BMI subgroups tested, suggesting that these models are appropriate to apply across a diverse adult population. Additionally, the weighted kappa scores were substantially higher than the unweighted kappa scores for our modeling methods, suggesting many "near misses" when predicting activity intensity. This would suggest that accelerometer data gives a lot, but not all, of the information necessary to correctly assess physical activity intensity across all activity types. As indicated above, additional sensors to assess other physiological or movement variables may offer additional, unique information to assist with improvement of physical activity intensity assessment from body-worn devices. Our study had many notable strengths. The inclusion of a diverse adult population facilitated development of models that achieved similarly high agreement with criterion data across all tested subgroup analyses. Additionally, the use of multiple accelerometers and modeling methods enabled for critical analysis of multiple important questions regarding accelerometer use simultaneously. Finally, our use of a true independent sample for cross-validation gave better insight into expected accuracy of these models in a new population, whereas holdout approaches such as leave-one-out cross-validation, which is commonly used with small datasets, may overestimate accuracy of developed models when applied in a new setting (Montoye *et al.*, 2018).

Our study also had several limitations. Our use of direct observation as a criterion method suggested that all activities of a certain type resulted in a similar activity intensity, whereas a method such as indirect calorimetry would allow physiologic assessment of participant effort and intensity. However, our data included a large proportion of non-steady-state movements, which impose significant hurdles to indirect calorimetry analysis. Additionally, similar direct observation systems have been validated by other researchers and used in both laboratory and field-based settings (Lyden *et al.*, 2014; Marcotte *et al.*, 2019; Alhassan *et al.*, 2017) justifying our use as a criterion measure in this study. Additionally, while past studies have evaluated contributions of different feature sets to accuracy of machine learning models, our study chose a single feature set to constrain our analysis to a more reasonable amount of data; future analyses should continue to assess interrelationships among feature sets, modeling methods, and accelerometer placements to optimize physical activity intensity assessment. Another modeling issue is that we considered recursive partition based classification methods that are shown to result in biased selection of features (Loh, 2002). Future research will explore recently developed classification algorithms based on unbiased feature selection criteria such as, Subgroup Identification based on Differential Effect Search (SIDES) (Lipkovich *et al.*, 2011) and the Generalized Unbiased Interaction Detection and Estimation (GUIDE) (Loh *et al.*, 2015). It remains to explore the extent to which these computationally intensive algorithms affect the accuracy in predicting physical activity intensity levels. Finally, participants in this study were all healthy and capable of participating in vigorous-intensity exercise, which may make our sample more fit than the average adult population. Cross-validation of our models in a lower-fitness population is therefore required before use in such a population. Additionally, while absolute MET thresholds for determining activity intensity are appropriate for younger adults of average/above average fitness, thresholds scaled to fitness level are more appropriate for extremely high- or low-fitness individuals. It may be that our modeling approach could be made more accurate if individualized to fitness level, as has been attempted with other accelerometer analysis methods (Ozemek *et al.*, 2013).

In conclusion, our study found that an accelerometer worn on the ankle, coupled with ensemble machine learning methods, achieved optimal accuracy for assessment of physical activity intensity. A two-accelerometer ankle-left wrist accelerometer combination yielded minor improvements over the ankle alone, but there was no additional benefit from more accelerometers or other two-accelerometer combinations. Future research should investigate using additional physiologic or movement sensors to further improve physical activity intensity assessment.

Declarations

Funding: This project was supported by ASPiRE and CAST internal grants from Ball State University.
Conflict of interest: We have no conflicts of interest to disclose.
Ethical approval: All study procedures were approved by the Ball State University Institutional Review Board.

References

Ainsworth BE, Haskell WL, Herrmann SD, Meckes N, Bassett Jr DR, Tudor-Locke C, Greer JL, Vezina J, Whitt-Glover MC, Leon AS (2011). "2011 Compendium of Physical Activities: a second update of codes and MET values." *Medicine & Science in Sports & Exercise*, **43**(8), 1575–1581. doi:http://dx.doi.org/10.1249/MSS.0b013e31821ece12.

Alhassan S, Sirard JR, Kurdziel LB, Merrigan S, Greever C, Spencer RM (2017). "Cross-validation of two accelerometers for assessment of physical activity and sedentary time in preschool children." *Pediatric Exercise Science*, **29**(2), 268–277. doi:http://dx.doi.org/10.1123/pes.2016-0074.

Bangdiwala SI (2017). "Graphical aids for visualizing and interpreting patterns in departures from agreement in ordinal categorical observer agreement data." *Journal of Biopharmaceutical Statistics*, **27**(5), 773–783. doi:http://dx.doi.org/10.1080/10543406.2016.1273941.

Breiman L (1996). "Bagging predictors." *Machine Learning*, **24**(2), 123–140. doi:http://dx.doi.org/10.1007/BF00058655.

Breiman L (2001). "Random forests." *Machine Learning*, **45**(1), 5–32.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA. ISBN 0-534-98053-8; 0-534-98054-6. doi:http://dx.doi.org/10.1201/9781315139470.

Chowdhury AK, Tjondronegoro D, Chandran V, Trost SG (2017). "Physical activity recognition using posterior-adapted class-based fusion of multiaccelerometer data." *IEEE Journal of Biomedical and Health Informatics*, **22**(3), 678–685. doi:http://dx.doi.org/10.1109/JBHI.2017.2705036.

Cohen J (1968). "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." *Psychological Bulletin*, **70**(4), 213. doi:http://dx.doi.org/10.1037/h0026256.

Dannecker KL, Sazonova NA, Melanson EL, Sazonov ES, Browning RC (2013). "A comparison of energy expenditure estimation of several physical activity monitors." *Medicine & Science in Sports & Exercise*, **45**(11), 2105. doi:http://dx.doi.org/10.1249/MSS.0b013e318299d2eb.

Donaldson SC, Montoye AH, Imboden MT, Kaminsky LA (2016). "Variability of objectively measured sedentary behavior." *Medicine & Science in Sports & Exercise*, **48**(4), 755. doi:http://dx.doi.org/10.1249/MSS.0000000000000828.

Dong B, Montoye A, Moore R, Pfeiffer K, Biswas S (2013). "Energy-aware activity classification using wearable sensor networks." In *Sensing Technologies for Global Health, Military Medicine, and Environmental Monitoring III*, volume 8723, p. 87230Y. International Society for Optics and Photonics. doi:http://dx.doi.org/10.1117/12.2018134.

Ekelund U, Steene-Johannessen J, Brown WJ, Fagerland MW, Owen N, Powell KE, Bauman A, Lee IM, Series LPA, Group LSBW, *et al.* (2016). "Does physical activity attenuate, or even eliminate, the detrimental association of sitting time with mortality? A harmonised meta-analysis of data from more than 1 million men and women." *The Lancet*, **388**(10051), 1302–1310. doi:http://dx.doi.org/10.1016/S0140-6736(16)30370-1.

Farrahi V, Niemelä M, Kangas M, Korpelainen R, Jämsä T (2019). "Calibration and validation of accelerometer-based activity monitors: A systematic review of machine-learning approaches." *Gait Posture*, **68**, 285–299. doi:http://dx.doi.org/10.1016/j.gaitpost.2018.12.003.

Freedson PS, Melanson E, Sirard J (1998). "Calibration of the Computer Science and Applications, Inc. accelerometer." *Medicine & Science in Sports & Exercise*, **30**(5), 777–781. doi:http://dx.doi.org/10.1097/00005768-199805000-00021.

Freund Y, Schapire RE (1997). "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of Computer and System Sciences*, **55**(1), 119–139. doi:http://dx.doi.org/10.1006/jcss.1997.1504.

Hager ER, Treuth MS, Gormely C, Epps L, Snitker S, Black MM (2015). "Ankle accelerometry for assessing physical activity among adolescent girls: threshold determination, validity, reliability, and feasibility." *Research Quarterly for Exercise and Sport*, **86**(4), 397–405. doi:http://dx.doi.org/10.1080/02701367.2015.1063574.

Hibbing PR, Lamunion SR, Kaplan AS, Crouter SE (2018). "Estimating Energy Expenditure with ActiGraph GT9X Inertial Measurement Unit." *Medicine & Science in Sports & Exercise*, **50**(5), 1093–1102. doi:http://dx.doi.org/10.1249/MSS.0000000000001532.

Hogg RV, McKean J, Craig AT (2005). *Introduction to mathematical statistics*. Pearson Education. doi:https://doi.org/10.1080/10543406.2013.756334.

John D, Freedson P (2012). "ActiGraph and Actical physical activity monitors: a peek under the hood." *Medicine & Science in Sports & Exercise*, **44**(1 Suppl 1), S86. doi:http://dx.doi.org/10.1249/MSS.0b013e3182399f5e.

Kaminsky LA, Brubaker PH, Guazzi M, Lavie CJ, Montoye AH, Sanderson BK, Savage PD (2016). "Assessing physical activity as a core component in cardiac rehabilitation: a position statement of the american association of cardiovascular and pulmonary rehabilitation." *Journal of cardiopulmonary rehabilitation and prevention*, **36**(4), 217–229. doi:http://dx.doi.org/10.1097/HCR.0000000000000191.

Kerr J, Patterson RE, Ellis K, Godbole S, Johnson E, Lanckriet G, Staudenmayer J (2016). "Objective assessment of physical activity: classifiers for public health." *Medicine & Science in Sports & Exercise*, **48**(5), 951. doi:http://dx.doi.org/10.1249/MSS.0000000000000841.

Kim H, Loh WY (2001). "Classification trees with unbiased multiway splits." *Journal of the American Statistical Association*, **96**(454), 589–604. doi:http://dx.doi.org/10.1198/016214501753168271.

Landis JR, Koch GG (1977). "The measurement of observer agreement for categorical data." *Biometrics*, pp. 159–174. doi:http://dx.doi.org/10.2307/2529310.

Lipkovich I, Dmitrienko A, Denne J, Enas G (2011). "Subgroup identification based on differential effect searchâĂŤa recursive partitioning method for establishing response to treatment in patient subpopulations." *Statistics in Medicine*, **30**(21), 2601–2621. doi:http://dx.doi.org/10.1002/sim.4289.

Löf M, Henriksson H, Forsum E (2013). "Evaluations of Actiheart, IDEEA® and RT3 monitors for estimating activity energy expenditure in free-living women." *Journal of Nutritional Science*, **2**.

Loh WY (2002). "Regression tress with unbiased variable selection and interaction detection." *Statistica Sinica*, pp. 361–386.

Loh WY, He X, Man M (2015). "A regression tree approach to identifying subgroups with differential treatment effects." *Statistics in Medicine*, **34**(11), 1818–1833. doi:http://dx.doi.org/10.1002/sim.6454.

Lu K, Yang L, Seoane F, Abtahi F, Forsman M, Lindecrantz K (2018). "Fusion of Heart Rate, Respiration and Motion Measurements from a Wearable Sensor System to Enhance Energy Expenditure Estimation." *Sensors*, **18**(9), 3092. doi:http://dx.doi.org/10.3390/s18093092.

Lyden K, Petruski N, Mix S, Staudenmayer J, Freedson P (2014). "Direct observation is a valid criterion for estimating physical activity and sedentary behavior." *Journal of Physical Activity and Health*, **11**(4), 860–863. doi:http://dx.doi.org/10.1123/jpah.2012-0290.

Mackintosh K, Montoye AH, Pfeiffer K, McNarry M (2016). "Investigating optimal accelerometer placement for energy expenditure prediction in children using a machine learning approach." *Physiological Measurement*, **37**(10), 1728. doi:http://dx.doi.org/10.1088/0967-3334/37/10/1728.

Marcotte RT, Petrucci JG, Cox MF, Freedson PS, Staudenmayer JW, Sirard JR (2019). "Estimating Sedentary Time from a Hip-and Wrist-worn Accelerometer." *Medicine & Science in Sports & Exercise*. doi:http://dx.doi.org/10.1249/MSS.0000000000002099.

Matthews CE, Chen KY, Freedson PS, Buchowski MS, Beech BM, Pate RR, Troiano RP (2008). "Amount of time spent in sedentary behaviors in the United States, 2003–2004." *American Journal of Epidemiology*, **167**(7), 875–881. doi:http://dx.doi.org/10.1093/aje/kwm390.

Montoye AH, Begum M, Henning Z, Pfeiffer KA (2017). "Comparison of linear and nonlinear models for predicting energy expenditure from raw accelerometer data." *Physiological Measurement*, **38**(2), 343. doi:http://dx.doi.org/10.1088/1361-6579/38/2/343.

Montoye AH, Mudd LM, Biswas S, Pfeiffer KA (2015). "Energy Expenditure Prediction Using Raw Accelerometer Data in Simulated Free Living." *Medicine & Science in Sports & Exercise*, **47**(8), 1735–1746. doi:http://dx.doi.org/10.1249/MSS.0000000000000597.

Montoye AH, Pivarnik JM, Mudd LM, Biswas S, Pfeiffer KA (2016). "Validation and comparison of accelerometers worn on the hip, thigh, and wrists for measuring physical activity and sedentary behavior." *AIMS Public Health*, **3**(2), 298.

Montoye AH, Westgate BS, Fonley MR, Pfeiffer KA (2018). "Cross-validation and out-of-sample testing of physical activity intensity predictions with a wrist-worn accelerometer." *Journal of Applied Physiology*, **124**(5), 1284–1293. doi:http://dx.doi.org/10.1152/japplphysiol.00760.2017.

O'Driscoll R, Turicchi J, Beaulieu K, Scott S, Matu J, Deighton K, Finlayson G, Stubbs J (2018). "How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies." *British Journal of Sports Medicine*, pp. bjsports–2018. doi:http://dx.doi.org/10.1017/S0029665118001532.

Ozemek C, Cochran HL, Strath SJ, Byun W, Kaminsky LA (2013). "Estimating relative intensity using individualized accelerometer cutpoints: the importance of fitness level." *BMC Medical Research Methodology*, **13**(1), 53. doi:http://dx.doi.org/10.1186/1471-2288-13-53.

Pate RR, Pratt M, Blair SN, Haskell WL, Macera CA, Bouchard C, Buchner D, Ettinger W, Heath GW, King AC, *et al.* (1995). "Physical activity and public health: a recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine." *JAMA*, **273**(5), 402–407. doi:http://dx.doi.org/10.1001/jama.273.5.402.

Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, George SM, Olson RD (2018). "The physical activity guidelines for Americans." *JAMA*, **320**(19), 2020–2028. doi:http://dx.doi.org/10.1001/jama.2018.14854.

Quinlan JR (1993). "C4.5: Programming for machine learning." *Morgan Kauffmann*, **38**, 48.

Sasaki JE, Hickey A, Staudenmayer J, John D, Kent JA, Freedson PS (2016). "Performance of activity classification algorithms in free-living older adults." *Medicine & Science in Sports & Exercise*, **48**(5), 941. doi:http://dx.doi.org/10.1249/MSS.0000000000000844.

Staudenmayer J, He S, Hickey A, Sasaki J, Freedson P (2015). "Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements." *Journal of Applied Physiology*, **119**(4), 396–403. doi:http://dx.doi.org/10.1152/japplphysiol.00026.2015.

Toth LP, Park S, Springer CM, Feyerabend MD, Steeves JA, Bassett DR (2018). "Video-recorded validation of wearable step counters under free-living conditions." *Medicine & Science in Sports & Exercise*, **50**(6), 1315–1322. doi:http://dx.doi.org/10.1249/01.mss.0000535946.47131.ae.

Tremblay MS, Aubert S, Barnes JD, Saunders TJ, Carson V, Latimer-Cheung AE, Chastin SF, Altenburg TM, Chinapaw MJ (2017). "Sedentary behavior research network (SBRN)–terminology consensus project process and outcome." *International Journal of Behavioral Nutrition and Physical Activity*, **14**(1), 75. doi:http://dx.doi.org/10.1186/s12966-017-0525-8.

Troiano RP, McClain JJ, Brychta RJ, Chen KY (2014). "Evolution of accelerometer methods for physical activity research." *British Journal of Sports Medicine*, **48**(13), 1019–1023. doi:http://dx.doi.org/10.1136/bjsports-2014-093546.