International Journal of EXERCISE SCIENCE

Original Research

Ordinal Statistical Models of Physical Activity Levels from Accelerometer Data

SHAFAYET S. HOSSAIN†, DREW M. LAZAR‡, and MUNNI BEGUM‡

Department of Mathematical Sciences, Ball State University, Muncie, IN, USA

†Denotes graduate student author, ‡Denotes professional author

ABSTRACT

International Journal of Exercise Science 14(7): 338-357, 2021. Improvements in accelerometer technology has led to new types of data on which more powerful predictive models can be built to assess physical activity. This paper explains and implements ordinal random forest and partial proportional odds models which both take into account the ordinality of responses given explanatory accelerometer data. The data analyzed comes from 28 adults performing activities of daily living in two visits while wearing accelerometers on the ankle, hip, right and left wrist. The first visit provided training data and the second testing data so that an independent sample, cross-validation approach could be used. We found that ordinal random forest produces similar accuracy rates and better linearly weighted kappa values than random forest. On the testing set, the ankle produced the best accuracy rates (33.3%), followed by the left wrist (34.7%), hip (36.9%) and then the right wrist (37.3%) using the best performing decision model for a four-activity level response. Linearly weighted kappa values indicated substantial agreement. For a two-activity level response, the error rates on the ankle, hip, left wrist and right wrist were 15.5%, 15.9%, 16.5% and 18.8%, respectively. The partial proportional odds model had significant goodness of fit ($p < 0.0001$) and provided interpretable coefficients (at $p = 0.05$), but there was significant variability in accuracy. These models can be used on accelerometer data collected during exercise studies and levels of activity can be assessed without direct observation. This work also can lead to theoretical improvements of current modeling techniques that are used for this purpose.

KEY WORDS: Physiology, machine learning, classification, ordinal forest, regression trees

INTRODUCTION

A key component of physiology is being able to objectively measure the level of physical activity of individuals engaged in ordinary tasks and exercise (42). Accelerometers, devices that measure acceleration forces in different directions, have been used for this purpose since the 1980s when the first accelerometer-based physical activity monitor was developed. Initially, manufacturer specific activity "counts" were used to estimate physical activity through "cut points" which are counts per minute thresholds. Unfortunately, studies have found that this approach is overly specific to population, activity, and brand of accelerometer (44).

Since then the technology of these accelerometer-based physical activity monitors (accelerometers) have improved considerably in storage, battery capacity, size, and sensitivity

(44). Along with this, researchers have been able to access and utilize raw acceleration signal data for making estimation of physical activity responses rather than relying on count thresholds (7). This raw data is vast as it comprises acceleration signals along x, y, and z axes at 30-60 samples per second. Thus, significant computational challenges exist in efficiently and effectively extracting physical activity information from raw acceleration signal data.

An approach that has been taken is to summarize raw acceleration data at every 30-second epochs (29). These summaries, which comprise the new explanatory data, consist of summaries for each axis (mean, variance, minimum, maximum, and upper percentiles) and summaries of pair-wise correlations between these axes. This new data, along with ordinal activity level responses, can be used to build statistical or decision models. These models can then be used to make a prediction of the type of activity an accelerometer wearer is engaged in at any particular epoch. Types of models that have been used for this purpose include decision trees or ensembles of decision trees, such as random forest, boosting and bagging (16), and parametric linear and nonlinear models, among others (27). The predictive power of these models is tested either using data set aside from the training set or by using a testing set. The training set is the data used to build the predictive model. Using data set aside from the training set naturally leads to inflated accuracies despite efforts to counteract this effect through methods such as leave-one out, k-fold cross-validation, or out-of-bag validation (48). A testing set is data on the same response and explanatory variables as the training set but it is collected separately and "out of sample" and can give a better idea of the true or "real world" accuracy and quality of a predictive model.

The response variable with respect to explanatory accelerometer data is typically created by measurement by a separate device or observation in order to match data from the accelerometer with activity intensity as an ordinal variable. Data is collected from the accelerometer and matched with observations of activity intensity and the levels of the response in order are sedentary, light, moderate and vigorous (SED, LPA, MPA, VPA, respectively). In this paper, activity intensities were set in four categories as: ≤ 1.5 METs as sedentary (SED), 1.6-2.9 METs as light (LPA), 3.0-5.9 METs as moderate (MPA), and ≥6 as vigorous (VPA), where METs are metabolic equivalents as provided by the 2011 Compendium of Physical Activities (1). If a response variable is categorical, but of ordinal nature, then it is important to take this into account when building a model (36). For example, if a researcher makes errors in classifying a sedentary activity (SED) as moderate (MPA) as opposed to classifying a sedentary activity (SED) as light (LPA) then the researcher may be more likely to associate particular health outcomes with activity that is significantly more vigorous than the actual activity level a subject is engaged in. This, in turn, can lead to more mistaken conclusions and recommendations based on data analysis. Other examples, and a discussion of the importance of taking into account the ordinality of responses is given by Janitza, et al. (14). This paper makes a unique contribution to the literature in explaining, building and testing two types of models that take into account ordinal responses given explanatory accelerometer data. These two models are ordinal random forest, which is a non-parametric decision model, and partial proportional odds, which is a parametric generalized linear model.

The unique approach in this paper can be used as a basis for further development and improvement of models that make predictions about activity levels from accelerometer data but that don't account for ordinality of responses (16, 20, 27). While both models developed and applied here take into account the ordinal nature of responses, the partial proportional odds model provides interpretability of the effect of changes in explanatory variables. For example, we estimate the change in odds that a participant will be at most at any activity level per unit change in BMI and age and in change in sex. Establishing relationships between demographic and other explanatory variables and activity levels in this manner is of interest to physiologists and exercise scientists as physical activity levels are important determinants of health (47).

METHODS

Participants

The study which produced the accelerometer data that we analyzed in this paper had 30 participants. Each of these participants made two separate visits to the Ball State Clinical Exercise Physiology Laboratory. These visits were designed to produce training and testing sets, respectively, for model building and testing.

All of the 30 participants had no orthopedic limitations. Ten adults (n = 5 female) were chosen from each of the three age categories 18-39, 40-59 and 60-79 years. There is well-established variability in activity levels among these age groups and this distribution of participants allows models to be built for the general adult population (6, 39). Accordingly, in our training data, the distribution of the 1408 observations in the SED class was 27.8% in 18-39, 34.9% in 40-59 and 37.3% in 60-79, the distribution of the 1369 observations in the LPA class was 36.5% in 18-39, 31.1% in 40-59 and 32.4% in 60-79, the distribution of the 1080 observations in the LPA class was 31.1% in 18-39, 30.8% in 40-59 and 38.1% in 60-79 and the distribution of the 298 observations in the VPA class was 50% in 18-39, 37.6% in 40-59 and 12.4% in 60-79. Overall, by a chi-squared test, in the training data the relationship between age group and activity level was significant with a $p \approx 0$. Also, after accounting for covariates, all of our partial proportional odds model show increases in odds that a participant will be below an activity level with increases in age (Tables 5 and 6). Including this range of age groups allows our trained model to take this into account so that it can be applied without regard to age.

Table 1. Demographic information of the participants given as mean ± standard deviation.

	Total Sample	Male (n = 14)	Female (n = 14)
Age (yrs)	48 ± 19.6	48.5 ± 19.83	47.6 ± 20.15
Weight (lbs.)	176 ± 34.72	194.96 ± 27.53	157.3 ± 31.32
Height (in.)	68.5 ± 3.52	71 ± 2.62	65.9 ± 2.3
BMI	26.4 ± 4.16	27.03 ± 3.07	25.8 ± 5.07

During data collection, two participants had invalid data, which resulted in the dataset of 28 individuals (n = 14 female) for our study. Table 1 shows demographic information of the participants.

Protocol

Accelerometer data was collected in two exercise laboratory visits. Visit 1 produced the training data for building our models and Visit 2 produced the test data for assessment. Visit 1 was highly supervised by research staff and designed so that an array of activities with varying intensities and speed would be available and the model would be fully trained. Participants performed eleven activities starting with lying on a padded table for ten minutes. Then, ten activities were assigned from Table 2 below. For each participant, two were chosen from the sedentary category, four were chosen from the lifestyle/chore activity category and four from the ambulatory/exercise category. These activities were chosen at random and in a manner such that all the activities within each of the categories were performed by about the same number of participants. Each activity was done for five minutes, with the order of the activities progressing from sedentary to lifestyle/chore to ambulatory/exercise. There was 1-2 minutes rest between activities. Participants were requested to perform the sedentary and the lifestyle/chore activities as they would in their day-to-day lives. For the ambulatory/exercise activities they were required to maintain a consistent speed and intensity.

In Visit 2, less structure was provided by research staff with participants engaging in activities in the way they would in their day-to-day lives. This approach has been previously used for creating testing data to evaluate the generalizability of models to free-living settings (40, 28). Participants engaged in sixteen activities, each done for two to fifteen minutes. The participants were told to choose four activities from the sedentary category, four from the lifestyle/chore category and four from the ambulatory/exercise category. As previous research shows that adults engage mostly in sedentary activity, participants were asked to engage in activities from the sedentary category for at least 40 minutes (26, 45). Unlike visit 1, participants in visit 2 were allowed, within a framework, to choose the time they spent doing activities, the activities to perform and the order in which to perform them.

Table 2. List of physical activities.

Sedentary	Lifestyle/Chore	Ambulatory/Exercise
Reading, using a computer, watching television, writing, playing cards.	Standing, dusting, making a bed, folding laundry, sweeping, vacuuming, simulated gardening, picking up items from the floor.	Slow/fast overground walking, treadmill walking, overground jogging, treadmill jogging, stationary cycling, ascending stairs, descending stairs.

During both visits, each participant wore four ActiGraph GT9X Link accelerometers, on the left and right wrists, over the right hip, and on the right ankle. The traditional placement, with count-based measurement of activity level, was on the hip but different placements including the ankle and the wrist have recently been more commonly used (30, 34, 41). These additional locations are chosen for better compliance, particularly when worn on the wrist, and for improved ability to measure certain activity metrics, including steps, when worn on the ankle (43, 44). Models built using data from accelerometers placed on the thigh have shown higher accuracy than hip- or wrist-worn accelerometers, however, less comfort and lower compliance with the thigh accelerometer placement has also been reported (20, 30). The accelerometers were initialized to capture acceleration data along x, y, and z axes at a rate of 60 samples per second.

Data was summarized in 30-second non-overlapping epochs with features such as the mean, variance, minimum, maximum and the 70th, 80th, and 90th percentile, and pair-wise correlations of axes over each epoch. Also, the demographic variables sex, age, height and weight of each participant were recorded in the data.

The 2011 Compendium of Physical Activities provided MET values for each activity. Activity intensities were set in four categories as: ≤ 1.5 METs as sedentary (SED), 1.6-2.9 METs as light (LPA), 3.0-5.9 METs as moderate (MPA), and ≥ 6 as vigorous (VPA) (1). We also did a sub-analysis in two categories with < 3.0 METs as SLPA and ≥ 3.0 METs as MVPA. In both visits, at the end of each activity, participants took one to two minutes of rest before starting the next activity. Two researcher staff members carefully noted activity start/stop and intensity of activities and came to agreement upon transitions between activities. These observed activities and their intensities according to the MET scale served as the ground truth for development of prediction models using accelerometer data. Once ground truth data were coded according to activity intensity, this data was integrated into the data set into correct 30-second epochs. All participants signed informed consent prior to participating and all study procedures were approved by the Ball State University Institutional Review Board. This research was carried out fully in accordance to the ethical standards of the International Journal of Exercise Science (33).

Statistical Analysis

The statistical analysis in this paper is of training and testing data which includes an ordinal response of physical intensity levels (categorized as SED, LPA, MPA and VPA) and a set of summary statistics of acceleration measures and demographical variables as input features. We had a total sample size in both the training and test data sets of n = 4313 observations. For each subject we had approximately 4313/28 ≈ 154 observations. Each observation was a summary of the 30 second epoch of accelerometer data with each subject observed for approximately 154 × 30 = 4260 seconds = 77 minutes. The n = 4313 observations provide a rich amount of variability in activity intensities with a varied population in terms of gender, age, height, and weight performing a range of activities as described in the Protocol section above.

Two of the three models used in this analysis take into account the ordinal nature of these responses which is novel for the analysis of accelerometer data in the literature. The models we consider are: random forest, ordinal forest (with two different "performance functions", equal and proportional) and partial proportional odds. These models are explained in the remainder of the Methods section below and are applied in the Results section. We first present decision trees which is a basic decision model often used as a building block in so-called ensemble decision models. Then we discuss random forest and modifications of random forest that account for ordinality, in particular, ordinal forest. We then present a parametric model, that is, a model that assumes some distribution of the response variable of interest up to some unknown parameters to be estimated. The model that we explain and then apply is the partial proportional odds model.

Non-Parametric Classification Trees: Decision trees and classification models built on decision trees are popular classification tools. Among the reasons for their popularity is that they don't

assume an underlying distribution (non-parametric) and that they can be readily trained and improved on new data. From a root node, decision trees sequentially create branches which split features (or explanatory variables) until terminal nodes are reached. The feature to split on at each step is determined by a weighted measure of impurity of the resulting nodes. A number of impurity measures of a node t exist but in R the default impurity measure is the Gini criterion

$$G_t = \sum_k \sum_{k \neq l} p_k p_l$$

where p_i is the proportion of responses in node t of class $i = 1, \ldots, c$ where c is the number of classes. Splitting typically stops before nodes are pure to avoid models that are overspecific or "overfit". A decision about an observation can then be made by following the branches by the splits to a terminal node.

Random Forest: Classification trees render marked interpretability as one can follow the path of any observation to a terminal node. However, they can have considerable variability on different partitions of the training set and produce suboptimal accuracy rates. In order to counteract this, many ensemble methods build and utilize many different trees from the data to create a decision model. Specifically, random forests take a large number of bootstrap samples (samples drawn with replacement of size equal to the data set) and build decision trees on each of the bootstrap samples. As the trees of the forest are being built on bootstrap samples, it is nearly guaranteed that they will all contain different data and thus "look" at the full data set from different perspectives. To increase differences among the trees further, for each tree and at each node, the data is split on a random set of size $\sqrt{p}$ of all the input variables where p is the total number of input variables in the data set. Finally, a decision about any observation is made by majority vote of the created decision trees in the random forest. Also, instead of setting aside part of the training data for assessing model quality, the data left out in each of the bootstrap samples can be used for this purpose and an average error rate can be computed on these samples. This is known as the out-of-bag error.

Ordinal Random Forest: Random forests have been used successfully for classification of activity level responses given explanatory accelerometer data accelerometer data as processed in the Protocol section (31). However, random forests do not account for the ordinality of the response variable (in this case, in order, SED, LPA, MPA and LPA). Modifications of the splitting criteria, exist, however, that do account for ordinality of responses, and have not heretofore been used for accelerometer data. One such modification uses the Generalized Gini criterion given as

$$GG_t = \sum_k \sum_{k \neq l} W_{kl} p_k p_l$$

as an alternative impurity measure of a node t (5). W_{kl} are weights that increase with the distance of k to l, and thus, for any two given classes and their proportions in a particular node, the impurity of the node increases with the distance of k to l.

In this paper, we use a recent classification method known as ordinal forest that takes advantage of regression forests (12). Regression forests are a regression technique for the prediction of continuous, quantitative responses that uses the structure of random forests (4). Quantitative responses have a higher level of measurement than ordinal responses and thus are implicitly ordinal. One could treat the ordinal classifications numerically (in our case, SED, LPA, MPA and VPA as 1, 2, 3 and 4, for example) and appeal to a regression forest but such classifications are arbitrary and this approach is shown to not influence the quality of predictions (12). Instead, ordinal forest proceeds as follows:

1. Choose a large number of random, heterogeneous partitions of [0, 1] by *J* intervals where *J* is the number of classes in the response.
2. Represent classes by the midpoints of respective, ordered intervals in the partition.
3. Build regression forests on these sets of midpoints for each partition.
4. The partitions with the smallest out-of-bag errors are summarized and the regression forest built on that summary is the resulting model from ordinal forest.

In computing out-of-bag errors in step 4 above, different performance functions can be chosen, each of which emphasize different objectives of the model.

1. The "equal" performance function treats each class the same regardless of class size.
2. The "proportional" performance function attempts to lower error rates from the larger classes at the expense of smaller ones and thus attempts to lower the overall error rate.
3. Classes of interest can be emphasized though custom weighting at the expense of accuracy on other classes.

Principal Components for Partial Proportional Odds Model: Principal component directions are an orthogonal coordinate system fit to the feature space with each coordinate successively accounting for the greatest possible share of the remaining variability in the features (15). For our partial proportion odds model we fit 17 principal component directions to our 24 summary statistics of acceleration data. We call the 24 variables in these PCA directions PCA1 – PCA17. Each of these 17 variables is a linear combination of the original 24. In this case, these directions account for over 99% of the variability of the original 24 variables. Also, adding directions one at a time, we noted that out-of-bag error rates on the training set did not decrease at the point we reached 17 directions. We also expressed height and weight as BMI. We did this for the following reasons:

1. There is a significant amount of collinearity in the 24 summary statistics and rank deficiency prevents a linear model from being created on the full set of explanatory variables. In order to separate out the variability of the data along non-correlated directions, we represented the data in principal component directions. This is a common approach for dealing with a collinearity problem (11, 18, 32). There is also collinearity between height and weight and we express them together as BMI.

2. The principal components provide parsimony in constructing our parametric generalized linear model with less parameters to estimate by likelihood estimation. You lose interpretability of coefficients in the fitted model in this way, but the variables before being expressed in terms of PCA directions are summary statistics, anyway. Expressing variables height and weight as BMI also provides more parsimony.

When computing measures of model performance, our testing set was expressed in the same principal component coordinate system fit to the training data.

Partial Proportional Odds Model: The response variable in our data set has four ordinal categories which were obtained through discretizing a continuous variable at pre-specified cutoff values as given in the Protocol section. The response categories in our dataset are SED, LPA, MPA and VPA. The ordinal nature of the response justifies the application of a proportional odds cumulative logit model to create a more parsimonious and powerful parametric model (36). The proportional odds model links the odds of being below a particular ordinal category to the odds of being above a particular ordinal category to a linear function of the explanatory variables (linear in the parameters) through a logit transformation. That is, assuming we have c levels, the proportional odds model assumes,

$$logit(P(Y \leq k|x)) = \log\left(\frac{P(Y \leq k|x)}{P(Y > k|x)}\right) = \alpha_k + x'\beta, k = 1,2,3, \dots, c-1$$

for explanatory variables $x = (x_1, \dots, x_p) \in R^{px1}$, parameters $\beta = (\beta_1, \dots, \beta_p) \in R^{px1}$ and where Y is the ordinal response. The aim is then to estimate the parameters through observation and mathematical optimization. From the above equation, we see that for each unit increase in x_i we have a change in the log-odds of being below category k of β_i for $i = 1, \dots, p$. Solving for cumulative probabilities in the equation above leads to

$$\pi_k = P(Y \leq k|x) = \frac{\exp(\alpha_k + x'\beta)}{1 + \exp(\alpha_k + x'\beta)}, k = 1,2,3, \dots, c-1.$$

After parameters are estimated as $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ we can make a prediction about the ordinal category to which an observation of x belongs. To do so we choose the k that maximizes the estimated probability that an observation belongs to a category, i.e., we maximize

$$\widehat{P(Y = k|x)} = \begin{cases} \widehat{\pi}_k, k = 1 \\ \widehat{\pi}_k - \widehat{\pi}_{k-1}, k = 2, \dots, c-1 \\ 1 - \widehat{\pi}_{k-1}, k = c. \end{cases}$$

where

$$\hat{\pi}_k = \frac{\exp(\hat{\alpha}_k + x'\hat{\beta})}{1 + \exp(\hat{\alpha}_k + x'\hat{\beta})}, k = 1,2,3, \dots, c-1.$$

We see that $\hat{\beta}_i$ is the estimated change in the log-odds that Y will be in class k or below for a unit change in x. This estimation holds for all $k = 1, 2, 3, \ldots, c-1$ (the probability that Y will be at most in class c is 1, so the odds are "infinite"). We also see that the logits are parallel, i.e., they have same slopes given by the values in β but the intercepts given by the $\alpha_k{}'s$ can vary. This is the parallel assumption of the proportional odds model and with this assumption a more parsimonious model can be built to avoid overfitting and to make estimation more reliable. We also might relax the parallel assumption by replacing some of the $\beta_i's$ with $\beta_{ik}{}'s$ which allows the effects of the covariates on log-odds to vary for different classes. This is known as the partial proportional odds model and this is the model we use for all placement locations (hip, ankle, right wrist and left wrist) for our data sets.

Measures of Performance: The measures of model performance we look at in this paper are the error rate, the out-of-bag error rate, the kappa value and the linearly weighted kappa value. The error rate is simply the proportion of observations that are misclassified by the model. We report the error rates on testing sets. An out-of-bag error is the average of error rates of observations left out of bootstrap samples in the random forest model where the classification of each of these observations is determined by trees built on the sets they were left out of. Kappa values compare observed versus expected accuracies to account for classification by chance (19). We look at kappa values for two classifications and linearly weighted kappa values for four class classifications. Weighted kappa values treat misclassifications further away from true classifications as more significant than misclassifications closer to true classifications (8). 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 linearly weighted kappa measure to do so, with movement from the true classification from one class to another penalized equally. A scale often used to interpret kappa values is given in Table 3 (19).

Table 3. Interpreting Kappa.

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

In Figure 1, we present a schematic that summarizes the process from the collection of data on the 30 subjects through to the testing of model performance after our models are built and is as follows:

- **A.** Same 30 participants in 2 visits, 28 with valid data.
- **B.** Explanatory data are summaries of acceleration data along x,y,z axes of 30-second epochs plus demographic variables.
- **C.** Build partial proportional odds model. Built on PCA directions plus demographic variables.
- **D.** Build decision models: random and ordinal forest.
- **E.** Apply decision models to testing data.
- **F.** Apply partial proportional odds model to testing data (expressed in terms of PCA directions).
- **G.** Model performance measures (error rates, kappa statistics) from applying models to testing data.

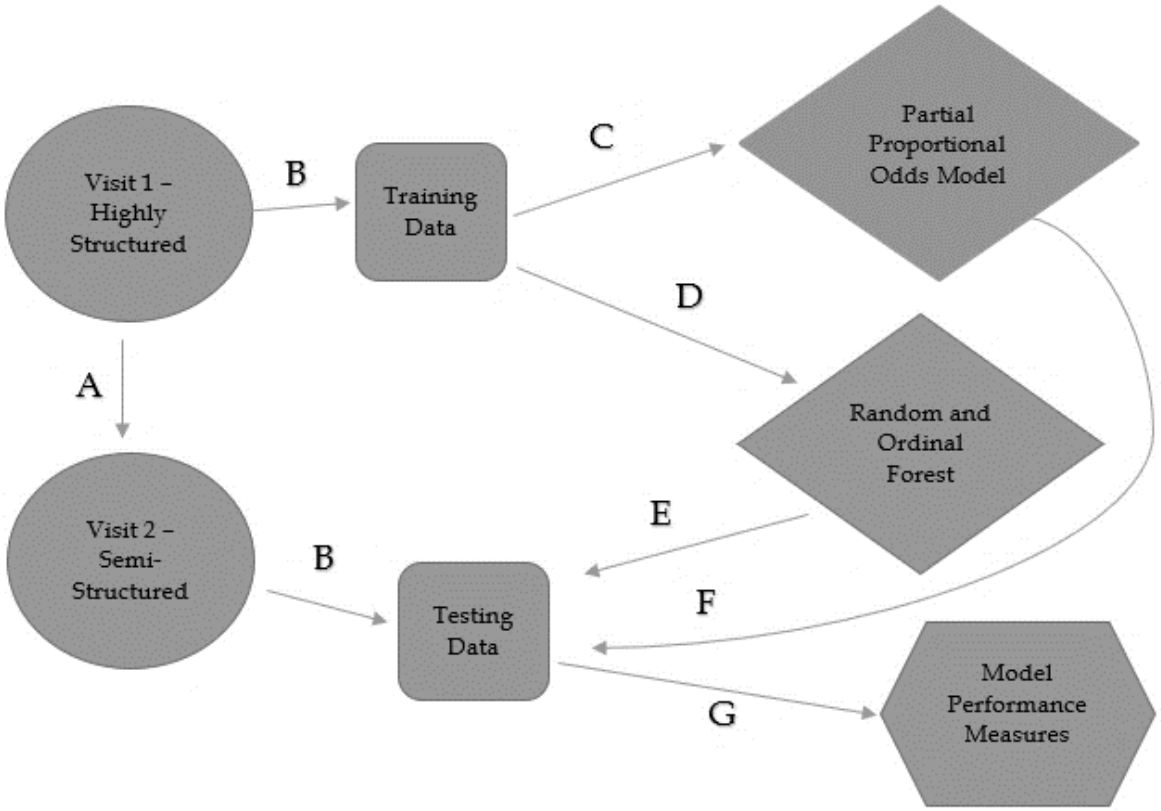

Figure 1. Schematic of collecting data through to testing model performance.

Ordinal forest models and random forests models were created using the R packages ordinal forest (13) and random forest (37), respectively. Partial proportional odds models were created in SAS 9.4 using proc logistic. Data from four participants in the study and all code used to create models, tables and figures in this paper is available at the following hyperlink, https://github.com/DrewLazar/Ord_Accelerometer.

RESULTS

We found that ordinal forest produces similar error rates but slightly better linearly weighted kappa values (a measure that takes into account both ordinality of responses and correct decisions by chance) than random forest overall. We also found that the ankle produces the best accuracy rates in our decision models and in our partial proportional odds model, but there was considerable variability in error rates across placements (ankle, hip, left and right wrist) in the partial proportional odds model. We summarize our analysis of our data sets using the decision models random forest and ordinal forest (with two different performance functions, equal and proportional). We then summarize our results using the partial proportional odds model and include tables of parameter estimates for the placements, the ankle and the hip, with the lowest error rates on the testing set.

As shown in Table 4, for all decision models, the ankle provides the lowest error rate on the testing set. It has a slightly higher out-of-bag error than the hip for random forest. The ankle also has the highest linearly weighted kappa for all decision models.

Table 4. Decision model performance for four classes.

Placement	LW	RW	HIP	ANK
	LWK/ER/*OBE	LWK/ER/OBE	LWK/ER/OBE	LWK/ER/OBE
Random Forest	0.6108/0.349/0.240	0.5702/0.376/0.233	0.608/0.367/**0.195**	0.652/0.333/0.208
Equal Ordinal Forest	0.6169/0.347	0.5723 /0.373	0.607/0.369	0.654/0.333
Proportional Ordinal Forest	0.6143/0.349	0.5738/0.372	0.609/0.367	**0.652/0.333**

LW=Left Wrist, RW=Right Wrist, ANK=Ankle, LWK=Linear Weighted Kappa, ER=Error Rate,
OBE = Out-of-Bag Error, *OBE for Random forest only, Lowest error rates and highest linearly weight kappa values in bold.

The ordinal forest models provide higher linearly weighted kappas than random forest for the left wrist and right wrist and linearly weighted kappas are nearly the same for the hip and the ankle. The results from proportional and equal ordinal forest are very similar on the testing set despite different weighting schemes according to class size used in training.

Figure 2 presents an agreement chart on the test set for the ankle accelerometer placement and using the proportional ordinal forest model (3). For each activity level, the width and height of the outer rectangle gives the marginal number of classifications of the activity level by the true classification and the classification algorithm, respectively. These widths and heights, in terms of number of classifications, are given along the top and right edge of the agreement plot. 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 classes. We can see in these plots that misclassifications outside of adjacent classes are relatively few. Other agreement plots for different placements look similar. This explains that the linearly weighted kappa values indicate substantial agreement as the linearly weighted kappa values account for ordinality of responses and the distance of classifications by the model to true classifications.

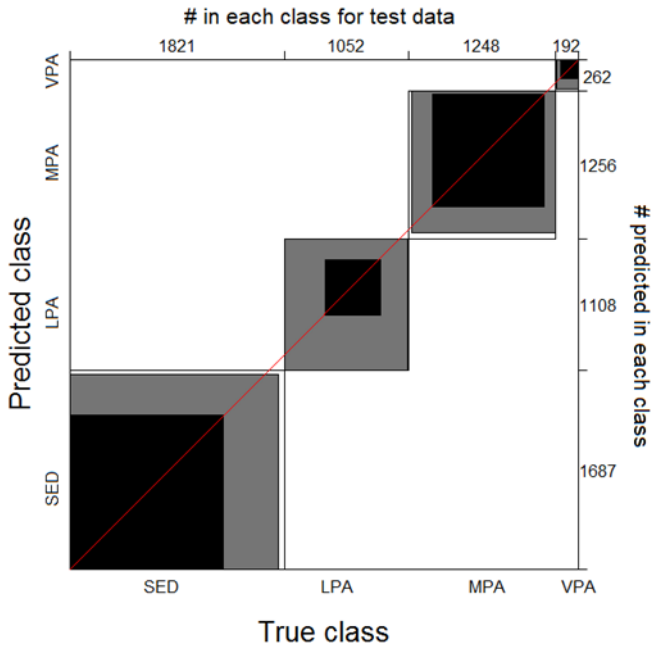

Figure 2. Agreement plot: Proportional ordinal forest for ankle.

With most misclassifications in adjacent classes, we collapsed our four category response into two classes, sedentary to light (SLPA), which includes SED and LPA and moderate to vigorous (MVPA) which include MPA and VPA and created random forest models from each of the accelerometer placements.

Table 5. Decision model performance for two classes.

Placement	LW	RW	HIP	ANK
	Kappa/ER/OBE	Kappa/ER/OBE	Kappa/ER/OBE	Kappa/ER/OBE
Random Forest	0.6165/0.165/0.128	0.5523/0.188/0.124	0.630/0.159/0.100	**0.742/0.115/0.040**

Lowest error rates and highest linearly weight kappa values in bold.

Accordingly, as shown in Table 5, we observed significantly lower error rates and out-of-bag error rates for our two-class model. Kappa values all indicate substantial agreement with the ankle providing the best summary statistics overall. Note that there must be more than two categories for the ordinality of responses to be meaningful and thus we do not use ordinal forest or the partial proportional hazards model when we collapse our four response classes (SED, LPA, MPA and VPA) into two classes (SLPA and MVPA).

Results for Partial Proportional Odds Model: We fit four partial proportional odds models for accelerometer placements on the left wrist, the right wrist, hip and the ankle. The results for the linearly weighted kappas and error rates are in Table 6. For each model we allowed parallel slopes for 17 principal component directions (PCA 1 – PCA 17) for parsimony and to ensure convergence in SAS.

Table 6. Partial Proportional log odds performance for four classes.

Placement	LW LWK/ER	RW LWK/ER	HIP LWK/ER	ANK LWK/ER
Proportional log odds	0.5435/0.403	0.4854/0.446	0.6596/0.329	**0.6714/0.318**

Lowest error rates and highest linearly weight kappa values in bold.

Every final model showed goodness of fit according to the likelihood ratio test (with the full model compared to the model fitted with just main effects) with p-values ≈ 0. For each placement we allowed slopes to vary for sex, age and BMI and we computed the AIC and BIC criteria and the error rate on the testing set. The AIC and BIC criteria are computed as AIC= $-2\ln L + 2p$ and BIC=$-2\ln L + \ln(n)p$ where n is the sample size and L is the likelihood of the model. As decreasing functions of L, and increasing functions of p, a smaller AIC or BIC is preferable (a smaller p means more parsimony).

For each placement, the model with the smallest BIC and the smallest error rate on the testing set agreed. The estimates for the models with the two lowest error rates, the ankle and the hip, are presented in Tables 7, 8a, and 8b, respectively. For the left wrist, right wrist, and ankle, where only age parameters varied by response level k = 1, 2, 3 the parameter estimates were very similar. Generally, the log-odds of being in a class or below increased with sex, age and BMI for these three models. For example, we can interpret $\hat{\beta}_{23}$ as the change in log-odds that a participant will be less than the moderate activity level for every unit increase in age given the other effects in the model. For the ankle, this is $\hat{\beta}_{23}$=0.039 so that the change in odds that a participant will be at most at the moderate activity level increases by a factor of exp (0.039) $\approx$ 1.04 or 4% for every unit increase in age. Thus, with all the age coefficients positive, probabilities that they will be below a particular activity level rise with age and older participants tends to be less active. A similar (but overall) effect for BMI, a 1 – exp (0.06) $\approx$ 6% increase in odds to be at most at any level below the vigorous category with unit increase in BMI, is observed. A positive overall effect for sex was also observed but it is not significant at the p = 0.05 significance level (but is significant for the left and right wrist models).

For the hip model, where sex is allowed to vary over the factor levels, the effects of the covariates are positive except for $\hat{\beta}_{22}$ which is not significant at the $p = 0.05$ significance level. Nearly all the parameters, in all the models, by the Wald test, are significant with $p < 0.05$.

As shown in Table 6, the error rates vary significantly among the models from 0.446 on the right wrist to 0.318 on the ankle which suggests the models could be more reliable, especially since our non-parametric decision models give similar error rates for different placements. The highest kappa values are for the ankle and in Figure 3 we present an agreement chart which shows, like for our decision models, that most classification errors are in adjacent classes.

Table 7. Ankle: Partial proportional log odds estimates.

Covariate	Intercept	Intercept	Intercept	Sex	Age	Age	Age	BMI
k	1	2	3	1,2,3	1	2	3	1,2,3
Parameter Estimate	$\hat{\alpha}_1 = -5.563$	$\hat{\alpha}_2 = -2.255$	$\hat{\alpha}_3 = 0.753$	$\hat{\beta}_1 = 0.110$	$\hat{\beta}_{21} = 0.014$	$\hat{\beta}_{22} = 0.006$	$\hat{\beta}_{23} = 0.039$	$\hat{\beta}_3 = 0.060$
Standard Error	0.335	0.324	0.370	0.072	0.002	0.003	0.005	0.009
p-value	< 0.0001	< 0.0001	0.0418	0. 1272	< 0.0001	0.028	< 0.0001	< 0.0001

Table 8a. Hip: Partial proportional log odds estimates.

Covariate	Intercept	Intercept	Intercept	Sex	Sex
k	1	2	3	1	2
Parameter Estimate	$\hat{\alpha}_1 = -5.655$	$\hat{\alpha}_2 = -2.05$	$\hat{\alpha}_3 = 0.693$	$\hat{\beta}_{11} = 0.045$	$\hat{\beta}_{12} = 0.218$
Standard Error	0.301	0.291	0.358	0.093	0.097
p	< 0.0001	< 0.0001	0.0525	0.591	0.0242

Table 8b. Hip: Partial proportional log odds estimates.

Covariate	Sex	Age	Age	Age	BMI
k	3	1	2	3	1,2,3
Parameter Estimate	$\hat{\beta}_{13} = 0.058$	$\hat{\beta}_{21} = 0.014$	$\hat{\beta}_{22} = 0.003$	$\hat{\beta}_{23} = 0.017$	$\hat{\beta}_3 = 0.008$
Standard Error	0.176	0.002	0.002	0.005	0.009
p	0.0009	< .0001	0.2811	0.0003	< .0001

DISCUSSION

In this paper, we explored decision and statistical models of data with explanatory accelerometer variables and an ordinal response variable that measured activity level (SED, LPA, MPA, VPA). Ordinal forest models provided similar error rates as random forest but somewhat better linearly weighted kappa values than random forest on two of four accelerometer placements. Partial proportional odds models gave results which varied considerably among the placements but produced a lower error rate and higher linearly weighted kappa value than any of the decision models on the ankle.

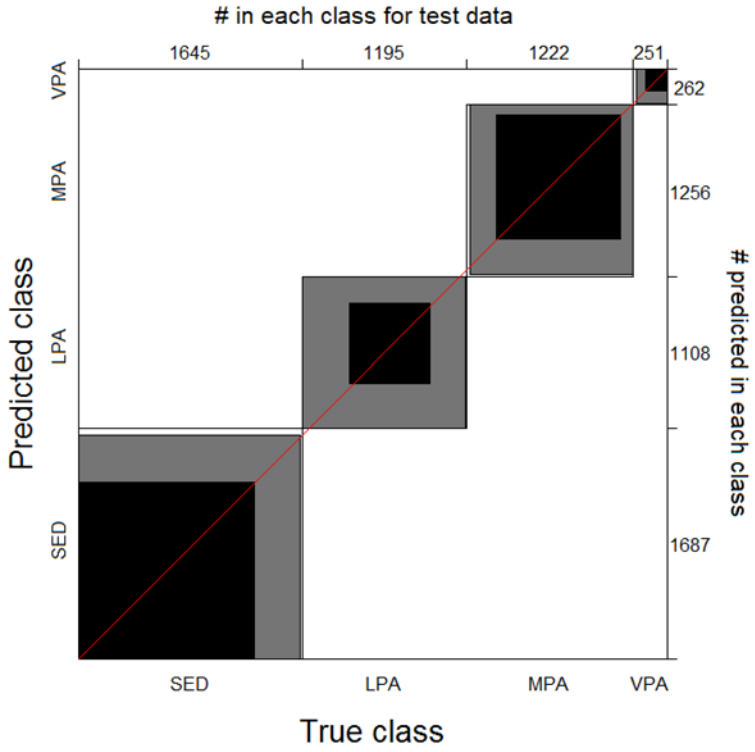

Figure 3. Agreement plot: Proportional log odds model for ankle.

Linearly weighted kappa values all indicated substantial agreement for all models except for the left wrist and right wrist in partial proportional odds which also had the highest error rates. Agreement plots for models with substantial agreement accordingly showed most misclassifications in adjacent classes. Also, for the random forest model, we collapsed four of our classes into two (SED and LPA to SLPA) and (MPA and VPA to MVPA) and found substantially lower error rates.

That the ankle provided the lowest error rates for the decision and partial proportional odds models agrees with other recent work that indicates that activity monitors worn somewhere on the lower limb provides better assessment of physical activity than devices worn on other body locations (9, 24). Other studies have shown that multiple accelerometer placements and sensors such as heart rate sensors or a gyroscope in addition to the accelerometer can increase accuracy in predicting physical activity intensity (10, 20, 23, 35).

Additional partial proportional odds models can be considered based on researcher interest and variable selection criteria. Generalized linear models can be challenging to fit, however, and with the large number of covariates present in the accelerometer data, additional dimension reduction techniques such as LASSO or feature selections and different ways to process the raw data in the initial step besides summary statistics can be explored.

Error rates on the testing set for four class prediction and using ordinal forest decision models, varied from 33.3% for the ankle placement to 37.6% for the left wrist placement. These are similar

to error rates found in other studies, some of which, however, use different activities, protocols, and model assessment. For example, our error rates are slightly better than those found by Sasaki et. al which uses a five-class prediction with a random forest model but worse than those found in Lazar et. al which compares combinations of multiple accelerometers placement as features in the model (20, 40). The use of linearly weighted kappa values, which not only takes into account correct assignment by chance but also penalizes misclassification according to the order of responses, isn't as common in these type of studies. Also, even though random forest methods naturally provide validation data when sampling from the training data, we did not find out-of-bag errors used in similar studies. Montoye et. al, however, used leave-one-out cross validation in random forest and found similar validation error rates (21.6% - 23.3%) that we found in our study (16).

Practically, accelerometers can be used to provide immediate, objective feedback to wearers about their level of physical activity in free living settings. This can be used to encourage increased physical activity which, in turn, is a crucial aspect of managing and improving health (46). Exercise scientists can use developed models to objectively classify levels of physical activity, to establish relationships between levels of physical activity and health indicators, and to develop guidelines for minimal or optimal levels of physical activity (17, 21, 38). However, Sasaki et. al notes overall error rates of classification algorithms of free-living physical activity from explanatory accelerometer data above the rate of 20% which they consider "acceptable" (40). This study provides an approach that can be extended and built upon to not only address these error rates but also take into the ordinality of responses in classification and misclassification. Accordingly, in addition to ordinal forest considered here, the weighted Gini index, GG_t, as an impurity measure or the twoing method in random forest, for example, can be coded in R (2). Similar to our results here, recent research on methods that account for ordinality of responses, however, note slightly better or similar error rates as ordinary decision models (e.g. random forest) (12, 14). In our activity level response, however, prior ordinal states inform later states, and with consideration of the ordinal nature of response, time series approaches can be used to build classification models on the training data and prior predictions can be used in testing and application of the models.

Our study has several notable strengths. Data derived from diverse, adult participants allowed the development of models that achieved substantial agreement with criterion responses. This is according to a linear weighted kappa statistic that takes into account the "distance" a misclassification is from the true classification by the order of the data. Our agreement plots, in Figures 2 and 3, reflect this substantial agreement provided by our models. Thus, our models and modeling methods can be applied to a general population with confidence that if there is misclassification then it is taking place mostly in adjacent classes. Both the assessment of our models, and, importantly, the development and application of ordinal forest and partial proportional odds models take this into account. Our results and this approach can be used as a basis for future model development that takes into the ordinality of responses. Also, the use of an independent sample for testing gives insight into the expected accuracy of these models in a new population, as opposed to validation methods which may overestimate accuracy of developed models when applied in a new setting (31).

Our study also had several limitations. The study that produced the data used direct observation to establish a criterion, which assumes that all activities of a certain type, as given by the 2011 Compendium of Physical Activities, result in the same activity level. A method such as indirect calorimetry would instead permit physiologic assessment of effort and activity levels. However, the number of "non-steady state movements" would make such assessment difficult (20). Regardless, similar direct observation systems have been used and validated by other researchers in the laboratory and in the field (22, 25). Also, while the study from which we analyzed our data tried to simulate a range of typical free-living daily activities in testing, there was not labeling of the data by specific activity which would allow stratification, model building and comparisons of accuracy rates for different placements according to different type of activity. This should be a subject of future research to allow exercise scientists and physiologists a choice of models and placements if they intend to assess physical activity levels from a specific type of prescribed activity. Also, other characteristics that might be useful, such as dominant hand of each participant, were not available but might be useful for analysis (although there was not a great difference in results for right and left wrists).

In addition, even though our parametric, partial proportional odds model produced the lowest error rates and highest linearly weighted kappa values for the ankle placement (0.318 and 0.6714, respectively), the error rates were significantly higher and linearly weighted kappa values were significantly lower on the left wrist (0.5435 and 0.403, respectively) and the right wrist (0.4854 and 0.446, respectively) for the same model. Parametric models are less robust then classification models built on decision trees, and thus these results for our partial proportional odds model could be specific to this data set with less generalizability than our random forest and ordinal forest decision models. This as well should be a subject of future research and labeling of specific activities in the data set might help discern why our partial proportional odds model produced the lowest error rates on the ankle in testing. Ordinality of response variables is an important consideration in classification problems. In addition to the methods used in this paper, there is room for different approaches and procedures which take advantage of this essential property of responses for better prediction and classification of physical activity levels from explanatory accelerometer data.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Alex Montoye for providing us with the accelerometer data. His study that generated the original accelerometer data was supported by ASPiRE and CAST internal grant from Ball State University.

REFERENCES

1. Ainsworth BE, Haskell WL, Herrmann SD, Meckes N, Bassett DR Jr, Tudor-Locke C, et al. 2011 Compendium of physical activities: a second update of codes and MET values. Med Sci Sports Exerc 43(8): 1575-1581, 2011.

2. Archer KJ. rpartOrdinal: An R package for deriving a classification tree for predicting an ordinal response. J Stat Softw 34: 7, 2010.

3. Bangdiwala SI, Shankar V. The agreement chart. BMC Med Res Methodol 13(1): 97, 2013.

4. Breiman L. Random forests. Mach Learn 45(1): 5-32, 2001.

5. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Boca Raton, FL: CRC press; 1984.

6. Caspersen CJ, Pereira MA, Curran KM. Changes in physical activity patterns in the United States, by sex and cross-sectional age. Med Sci Sports Exerc 32(9): 1601-9, 2000.

7. Chen KY, Bassett DR Jr. The technology of accelerometry-based activity monitors: current and future. Med Sci Sports Exerc 37(11 Suppl): S490-S500, 2005.

8. Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 70(4): 213-220, 1968.

9. Chowdhury AK, Tjondronegoro D, Chandran V, Trost SG. Physical activity recognition using posterior-adapted class-based fusion of multiaccelerometer data. IEEE J Biomed Health Inform 22(3): 678-685, 2018.

10. Dannecker KL, Sazonova NA, Melanson EL, Sazonov ES, Browning RC. A comparison of energy expenditure estimation of several physical activity monitors. Med Sci Sports Exerc 45(11): 2105-2112, 2013.

11. Dormann C, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36(1): 27-46, 2013.

12. Hornung, R. Ordinal forests. J Classification 37: 4-17, 2020.

13. Hornung, R. ordinalForest: Ordinal forests: prediction and variable ranking with ordinal target variables. R package version 2, 2018.

14. Janitza S, Tutz G, Boulesteix AL. Random forest for ordinal responses: prediction and variable selection. Comput Stat Data Anal 96: 57-73, 2016.

15. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans A Math Phys Eng Sci 374(2065): 20150202, 2016.

16. Kerr J, Ellis K, Godbole S, Staudenmayer J, Lanckriet G. Hip and wrist accelerometer algorithms for free-living behavior classification. Med Sci Sports Exerc 48: 933–940, 2016.

17. 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.

18. Lafi S, Kaneene J. An explanation of the use of principal-components analysis to detect and correct for multicollinearity. Prevent Vet Med 13(4): 261-275, 1992.

19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 33(1): 159-174, 1977.

20. Lazar D, Begum M, Murshed MM, Nelson B, Bock JM, Imboden M, et al. Statistical learning methods to predict activity intensity from body-worn accelerometers. J Biomed Anal 3(1): 27-50, 2020.

21. Lim WK, Davila S, Teo JX, Yang C, Pua CJ, Blöcker C, et al. Beyond fitness tracking: the use of consumer-grade wearable data from normal volunteers in cardiovascular and lipidomics research. PLoS Biol 16(2): e2004285, 2018.

22. 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(4): 860-3, 2014.

23. Lu K, Yang L, Seoane F, Abtahi F, Forsman M, Lindecrantz K. Fusion of heart rate, respiration and motion measurements from a wearable sensor system to enhance energy expenditure estimation. Sensors (Basel) 18(9): 3092, 2018.

24. Mannini A, Rosenberger M, Haskell WL, Sabatini AM, Intille SS. Activity recognition in youth using single accelerometer placed at wrist or ankle. Med Sci Sports Exerc 49(4): 801-812, 2017.

25. Marcotte RT, Petrucci GJ Jr, Cox MF, Freedson PS, Staudenmayer JW, Sirard JR. Estimating sedentary time from a hip- and wrist-worn accelerometer. Med Sci Sports Exerc 52(1): 225-232, 2020.

26. 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(7): 875-881, 2008.

27. Montoye AHK, Begum M, Henning Z, Pfeiffer KA. Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data. Physiol Meas 38(2): 343-357, 2017.

28. Montoye AH, Mudd LM, Biswas S, Pfeiffer KA. Energy expenditure prediction using raw accelerometer data in simulated free living. Med Sci Sports Exerc 47(8): 1735-1746, 2015.

29. Montoye AHK, Nelson MB, Bock JM, Imboden MT, Kaminsky LA, Mackintosh KA, et al. Raw and count data comparability of hip-worn ActiGraph GT3X+ and Link accelerometers. Med Sci Sports Exerc 50(5): 1103-1112, 2018.

30. Montoye AHK, 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(2): 298-312, 2016.

31. Montoye AHK, 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(5): 1284-1293, 2018.

32. Næs, T, Mevik B-H. Understanding the collinearity problem in regression and discriminant analysis. J Chemometrics 15(4): 413-426, 2001.

33. Navalta JW, Stone WJ, Lyons TS. Ethical issues relating to scientific discovery in exercise science. Int J Exerc Sci 12(1): 1-8, 2018.

34. Ozemek C, Kirschner MM, Wilkerson BS, Byun W, Kaminsky LA. Intermonitor reliability of the GT3X+ accelerometer at hip, wrist and ankle sites during activities of daily living. Physiol Meas 35(2): 129-38, 2014.

35. O'Driscoll R, Turicchi J, Beaulieu K, Scott S, Matu J, Deighton K, et al. How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. Br J Sports Med 54(6): 332-340, 2020.

36. Penn State University. STAT 504 Sec. 8.1 - Polytomous Logistic Regression. Retrieved from : https://online.stat.psu.edu/stat504/node/172/; 2020.

37. RColorBrewer S, Liaw MA. R Package 'randomForest.' Berkeley, CA: University of California; 2018.

38. Retner R. Fitness & big data: how wearable tech is changing exercise research. Retrieved from: https://www.livescience.com/45634-accelerometers-exercise-research.html; 2020.

39. Sallis JF. Age-related decline in physical activity: a synthesis of human and animal studies. Med Sci Sports Exerc 32(9): 1598-600, 2000.

40. 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(5): 941-950, 2016.

41. Strath SJ, Kate RJ, Keenan KG, Welch WA, Swartz AM. Ngram time series model to predict activity type and energy cost from wrist, hip and ankle accelerometers: implications of age. Physiol Meas 36(11): 2335-51, 2015.

42. Sylvia LG, Bernstein EE, Hubbard JL, Keating L, Anderson EJ. Practical guide to measuring physical activity. J Acad Nutr Diet 114(2): 199-208, 2014.

43. Toth LP, Park S, Springer CM, Feyerabend MD, Steeves JA, Bassett DR. Video-recorded validation of wearable step counters under free-living conditions. Med Sci Sports Exerc 50(6): 1315-1322, 2018.

44. Troiano RP, McClain JJ, Brychta RJ, Chen KY. Evolution of accelerometer methods for physical activity research. Br J Sports Med 48(13): 1019-1023, 2014.

45. United States. Bureau of Labor Statistics. American Time Use Survey. Retrieved from: https://www.bls.gov/tus/; 2020.

46. Ward DS, Evenson KR, Vaughn A, Rodgers AB, Troiano RP. Accelerometer use in physical activity: best practices and research recommendations. Med Sci Sports Exerc 37(11): S582-8, 2005.

47. Warburton D, Bredin S. Health benefits of physical activity: a systematic review of current systematic reviews. Curr Opin Cardiol 32(5): 541-556, 2017.

48. Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Patt Rectn 48(9): 2839-2846, 2015.