***Journal for the Measurement of Physical Behaviour*, 2019, 2, 237-246**
https://doi.org/10.1123/jmpb.2018-0011

Human Kinetics
ORIGINAL RESEARCH

Cross-Validation and Comparison of Energy Expenditure Prediction Models Using Count-Based and Raw Accelerometer Data in Youth

Alexander H.K. Montoye
Alma College

Kimberly A. Clevenger
Michigan State University

Kelly A. Mackintosh and Melitta A. McNarry
Swansea University

Karin A. Pfeiffer
Michigan State University

Background: Machine learning may improve energy expenditure (EE) prediction from body-worn accelerometers. However, machine learning models are rarely cross-validated in an independent sample, and the use of machine learning raises additional questions including the effect of accelerometer placement and data type (count vs. raw) for optimal EE prediction. **Purpose**: To assess the accuracy of artificial neural network (ANN) models for EE prediction in youth using count-based or raw data from accelerometers worn on the hip, wrist, or in combination, and compare these to count-based, EE regression equations. **Methods**: Data were collected in two settings; one ($n = 27$) to calibrate the EE prediction models, and the other ($n = 34$) for model cross-validation. Participants wore a portable metabolic analyzer (EE criterion) and accelerometers on the left wrist and right hip while completing 30 minutes of exergames (calibration, cross-validation) and a maximal exercise test (calibration only). Six ANNs were created from the calibration data, separately by accelerometer placement (hip, wrist, combination) and data format (count-based, raw) to predict EE (15-second epochs). Three count-based linear regression equations were also developed for comparison to the ANNs. **Results**: The count-based, hip ANN demonstrated lower error (RMSE: 1.2 METs) than all other ANNs (RMSE: 1.7–3.6 METs) and EE regression equations (RMSE: 1.5–3.2 METs). However, all models showed bias toward the mean. **Conclusion**: An ANN developed for hip-worn accelerometers had higher accuracy for EE prediction during an exergame session than wrist or combination ANNs, and ANNs developed using count-based data had higher accuracy than ANNs developed using raw data.

Keywords: activity trackers, machine learning, out-of-sample, pattern recognition, physical activity

Despite the well-known benefits of physical activity (PA) participation in youth, the majority do not meet recommended PA levels (Esteban-Cornejo, Tejero-Gonzalez, Sallis, & Veiga, 2015; Troiano et al., 2008; US Department of Health and Human Services, 2018). Measurement of energy expenditure (EE) using accelerometers is common for determining the volume and intensity of PA, and accurate EE measurement is critical for identification of, and intervention in, youth with low PA. Due to memory capacity and battery life limitations, early accelerometers summarized raw data into 'activity counts' or other condensed storage forms on-board the accelerometer in 1–60+ second intervals (epochs), meaning that raw data were not available for download. Newer accelerometers allow access to raw (*g*) data collected at high sampling rates for days or weeks at a time (John & Freedson, 2012).

EE prediction models developed for count-based accelerometer data are inherently limited in their applicability to other accelerometer brands because counts are brand-specific and often proprietary (John & Freedson, 2012). The use of raw data has the potential to improve the application of models across accelerometer brands and give transparency to features and models used to interpret accelerometer data and associated outcomes (van Hees et al., 2014; van Hees et al., 2013). However, the majority of studies that have developed models for EE prediction have relied on raw data metrics (e.g., mean and percentiles of signal) which are subject to orientation-dependency and are, therefore, influenced by factors such as the angle, attachment method, and side of the body on which an accelerometer is worn. Conversely, count-based data are non-negative and cumulative and, thus, less likely to be influenced by device orientation. Such differences in these types of data may make predictive models using orientation-dependent raw data more prone to over-fitting (Montoye, Pivarnik, Mudd, Biswas, & Pfeiffer, 2016). The vector magnitude (VM) of triaxial accelerometers has been proposed as a strategy to alleviate such issues of orientation dependency for both count and raw data and has been used with hip-worn accelerometers and accelerometers placed on alternate locations (Sasaki, John, & Freedson, 2011; van Hees et al., 2013). However, whether VM improves EE prediction accuracy compared to using triaxial data is equivocal, especially when data type (count vs. raw) and accelerometer placement are considered (Montoye et al., 2016).

Several types of predictive models have been developed for translating accelerometer data into EE in youth, ranging in complexity from count-based regression models (Crouter, Horton, &

Montoye is with Integrative Physiology and Health Science, Alma College, Alma, MI. Clevenger and Pfeiffer are with the Department of Kinesiology, Michigan State University, East Lansing, MI. Mackintosh and McNarry are with The Applied Sports Technology Exercise and Medicine Research Centre, Swansea University, Swansea, Wales, United Kingdom. Montoye (montoyeah@alma.edu) is corresponding author.

Bassett, 2012; Freedson, Pober, & Janz, 2005) to machine learning models (Mackintosh, Montoye, Pfeiffer, & McNarry, 2016; Trost, Wong, Pfeiffer, & Zheng, 2012), which use count-based or raw data as inputs. Machine learning models have generally yielded more accurate predictions of EE than linear regression models in initial calibration settings in both youth (Mackintosh et al., 2016; Trost et al., 2012) and adult samples (Montoye, Begum, Henning, & Pfeiffer, 2017). However, several studies in adults (Gyllensten & Bonomi, 2011; Lyden et al., 2014; Sasaki et al., 2016; Staudenmayer et al., 2015) and one study in youth (Hibbing, Ellingson, et al., 2018) have demonstrated that the accuracy of these models decreases when cross-validating in a new or independent sample, indicating a tendency for machine learning models to be over-fit to the data. Further research is, therefore, required to determine the accuracy of machine learning models for the prediction of EE in independent data sets.

Finally, it must be considered that the potential benefit of raw data and/or machine learning modeling for EE prediction may be dependent on number and placement of accelerometers used. While the hip is the most common accelerometer placement, wrist-worn accelerometers have seen increased use in recent years due to improved wear-time compliance and ability to capture behaviors such as activity type and sleep (Montoye, Moore, Bowles, Korycinski, & Pfeiffer, 2018; Troiano, McClain, Brychta, & Chen, 2014). Indeed, early research utilizing count-based data has shown poorer accuracy of wrist-worn, compared to hip-worn, accelerometers in adults (Bouten, Sauren, Verduin, & Janssen, 1997; Swartz et al., 2000). More recent studies in youth and adult samples indicate that EE prediction from wrist-worn accelerometers is improved when more complex modeling approaches and/or triaxial/VM data were used instead of vertical axis data and/or simple linear regression models on activity count data (Crouter, Flynn, & Bassett, 2015; Montoye et al., 2017; O'Driscoll et al., 2018). Conversely, EE prediction from hip-worn accelerometers may be less affected by modeling method (Montoye et al., 2017). Additionally, use of multiple accelerometers placed on different body locations sometimes, but not always, leads to improved prediction accuracy (Mackintosh et al., 2016). More research is needed to understand the accuracy of accelerometers worn on various body locations.

Given the current gaps in our understanding of whether data type (count vs. raw), accelerometer number (one vs. multiple), and modeling method (machine learning vs. linear regression) affect EE prediction accuracy in youth, the present study's aims were to use an independent sample, cross-validation design to 1) determine if count-based or raw data inputs into a machine learning model yield better EE prediction accuracy, 2) investigate whether hip- or wrist-worn accelerometer data as model inputs (or a combination thereof) yield higher EE prediction accuracy, and 3) compare the accuracy of these machine learning models to three count-based EE prediction regression equations.

Methods

In the present study, we describe the development (calibration) of six artificial neural networks (ANNs) and then focus on the EE prediction accuracy of these ANNs and three count-based regression models in an independent, cross-validation setting. Each institution's respective ethics board approved this study, and participants and parents/guardians provided assent and consent prior to completing the study, respectively.

Calibration Participants and Protocol

Descriptions of the sample and procedures used for the calibration portion of this study have been described previously (Mackintosh et al., 2016). Briefly, 27 youth (15 boys; 11.6 ± 1.0 years) from Swansea, UK participated in a protocol in which they played active video games (exergames; two sessions of 15 minutes with a break between sessions; games included River Rush and Kinect Adventures Reflex Ridge on Xbox 360) and an incremental, graded treadmill test to volitional exhaustion. During the protocol, participants wore an ActiGraph wGT3X-BT (ActiGraph Corp., Pensacola, FL) accelerometer at the right hip and left wrist (collecting raw, triaxial data at 100 Hz) and a METAMAX 3B (Biophysik, Leipzig, Germany) metabolic analyzer (collecting breath-by-breath oxygen consumption).

Cross-Validation Participants and Protocol

Participants in the cross-validation study were 34 youth (Table 1; 21 boys; 10.3 ± 1.1 years) from the community of East Lansing, MI. Participants were free from any metabolic or physical condition that would alter their ability to perform, or alter their metabolic response to, the study protocol.

Participants were asked to visit the laboratory for a single visit at least two hours postprandial and having avoided caffeine and strenuous exercise for at least 24 hours prior to their visit. Initially, stature was measured to the nearest 0.01 m (Harpenden stadiometer, Holtain, Crymych, United Kingdom) and body mass to the nearest 0.1 kg (Seca digital scale, Hamburg, Germany) using standardized procedures.

Subsequently, two exergames were selected at random from a list of four games previously shown to elicit moderate- to vigorous-intensity PA [MVPA; Kinect Adventures Reflex Ridge, Just Dance 3, Wipeout, and Kinect Sports Boxing; (Barkman, Pfeiffer, Diltz, & Peng, 2016; Clevenger & Howe, 2015; Rosenberg et al., 2013)]. The two games were completed on the easiest level in both single- and multi-player mode (with a research assistant or friend/sibling) for 15 minutes each, resulting in four conditions (two games × two modes). Between games, participants were provided with a 5- to 10-minute break (e.g., to drink water, use the restroom).

Table 1 Participant Characteristics in Calibration and Cross-Validation Samples

	Calibration			Cross-validation		
	Boys (*n* = 15)	Girls (n = 12)	Total (*N* = 27)	Boys (n = 21)	Girls (n = 13)	Total (*N* = 34)
Age (years)	10.8 (1.2)	10.8 (1.4)	10.8 (1.0)	11.6 (1.4)	11.5 (1.1)	11.6 (1.2)
Stature (m)	1.46 (0.13)	1.45 (0.10)	1.45 (0.11)	1.57 (0.14)	1.54 (0.11)	1.56 (0.12)
Mass (kg)	38.7 (8.5)	37.2 (9.0)	38.7 (8.8)	49.3 (14.2)	48.1 (15.1)	48.8 (14.6)
Predicted basal VO_2 ($ml \cdot kg^{-1} \cdot min^{-1}$)	4.9 (0.4)	4.6 (0.7)	4.8 (0.5)	4.4 (0.5)	4.0 (0.7)	4.3 (0.6)

Note. Data are shown as mean (standard deviation).

Throughout all sessions and breaks (with the exception of participants consuming water or using the restroom), breath-by-breath gas exchange was assessed (Oxycon Mobile, Carefusion, Yorba Linda, CA, USA). The Oxycon has been shown to provide reliable and accurate measures of oxygen consumption compared to the Douglas bag method (Rosdahl, Gullstrand, Salier-Eriksson, Johansson, & Schantz, 2010). Additionally, two ActiGraph GT3X+ accelerometers were worn, one on the right hip at the level of the anterior axillary line (orientation: y-axis vertical, x-axis medial-lateral, z-axis anterior-posterior) and one on the posterior aspect of their left wrist between the styloid processes of the radius and ulna (orientation: y-axis vertical, x-axis medial-lateral, z-axis anterior-posterior when in anatomical position); both were secured in place with elastic belts. The ActiGraph GT3X+ has a range of ± 6 gravitational (g) units, and all monitors were set to collect triaxial data in raw mode at a sampling rate of 30 Hz.

Data Processing

Data from the exergames, rest intervals, and for calibration, the incremental (graded) treadmill test, were used for this study. Thus, the calibration and cross-validation protocols included both steady-state and non-steady-state data.

For the calibration data only, the 100 Hz accelerometer data were downloaded in raw form and downsampled to 30 Hz to avoid issues in data comparability between calibration and cross-validation given previous work showing that the use of different sampling rates affects the conversion of raw data to activity counts (Brønd & Arvidsson, 2016). To downsample the data, the downloaded .gt3x files from the accelerometer were converted to .wav files using in-house Java software (Oracle Corp., Redwood Shores, CA). These files were subsequently read into MATLAB (MathWorks Inc., Natwick, MA) and resampled to 30 Hz using the *resample* function available in MATLAB (Lyons, 2013). Once resampled, the 30 Hz files were converted back to .gt3x files using the same Java program. All subsequent analyses were conducted with 30 Hz data.

For both calibration and cross-validation, six features (mean and variance from each of the three accelerometer axes) were calculated from the raw accelerometer data in 15-second epochs using the feature extraction tool in ActiLife version 6.13 software (ActiGraph Corp., Pensacola, FL, USA). Data were also downloaded as activity counts in 1-second epochs using the ActiLife software, and six features (mean and variance of the activity counts in each accelerometer axis) were calculated from this count data in 15-second epochs using Microsoft Excel 2013 (Microsoft Inc., Redmond, WA). These features were chosen in accord with previous research developing machine learning models to predict EE in youth (Mackintosh et al., 2016), and 15-second epochs were chosen as previous research has shown the transient activity patterns of youth may necessitate shorter epochs than the traditional 60-second epochs used in adults (Bailey et al., 1995).

Relative oxygen uptake data from the metabolic analyzers were downloaded in 15-second epochs and converted to corrected metabolic equivalents (METs). Specifically, equations adapted from Schofield (1985) were used to predict basal metabolic rate in kcals/day. Next basal metabolic rate was converted to milliliters of oxygen consumed per minute and subsequently to ml/kg/min for determination of age- and sex-specific youth metabolic equivalent (MET) values (FAO/WHO/UNU Expert Consultation, 2001; Schofield, 1985). For example, if a participant's basal metabolic rate was predicted to be 4.0 ml/kg/min, then Schofield-corrected METs were calculated for this participant by dividing their relative oxygen consumption data by 4.0 (rather than 3.5, which is common in adults). This procedure is supported by a position statement published by the CDC/NIC/NCCOR Research Group on Energy Expenditure in Children (McMurray et al., 2015).

Once accelerometer data features and criterion corrected METs were calculated, they were time-aligned. All 15-second epochs where criterion EE was <0.5 corrected METs were removed as this generally represents non-wear or poor sampling (e.g., occluded sample line) in a given epoch (Mackintosh et al., 2016); this resulted in removal of ~1.4% of the epochs in the calibration dataset and ~0.3% of the epochs in the cross-validation dataset.

Using the features calculated from the count-based and raw data, six ANNs were created (using calibration data) and then tested (using cross-validation data) to predict EE; these ANNs were hip count, wrist count, combination count, hip raw, wrist raw, and combination raw, wherein "combination" used a combination of both hip and wrist data. The ANNs included in this study were feedforward, had one hidden layer with five hidden units, and did not have skip-layer connections; all ANNs were developed using the *nnet* package in the R software (Ripley & Venables, 2016). Access to sample data and the ANNs can be found at the following link: https://drive.google.com/open?id=1SlnXJBh6WUpxJJAjAovVbNw8hW54PhbZ. ANNs were chosen instead of other machine learning models since previous research shows promise for their use in EE estimation from accelerometer data (Mackintosh et al., 2016; Montoye, Mudd, Biswas, & Pfeiffer, 2015; Preece et al., 2009; Staudenmayer, Pober, Crouter, Bassett, & Freedson, 2009).

Additionally, our study sought to compare our developed ANNs to traditional, regression-based EE prediction methods. However, no previous work has developed regression equations to predict EE from count-based, hip- or wrist-worn ActiGraph accelerometer data in 15-second epochs, and there are indications that epoch length affects accelerometer output (McClain, Abraham, Brusseau, & Tudor-Locke, 2008). Therefore, we developed (using calibration data) and tested (using cross-validation data) three in-house regression equations for predicting EE from the VM activity counts according to accelerometer placement (hip, wrist, or combination). These equations were developed in SPSS version 24.0 (IBM Corp., Armonk, NY, USA). The resulting equations are as follows, where "HVM" signifies VM counts ($VM = \sqrt{x^2 + y^2 + z^2}$) from the hip accelerometer per 15 seconds and "WVM" signifies VM counts from the wrist accelerometer per 15 seconds:

- E1: Hip: $\text{METs} = 0.002346 \times \text{HVM} + 2.576510$
- E2: Wrist: $\text{METs} = 0.000898 \times \text{WVM} + 2.495456$
- E3: Hip and wrist combination: $\text{METs} = 0.001078 \times \text{HVM} + 0.000591 \times \text{WVM} + 2.339118$

Cross-Validation Data Analysis

All data violated tests for normality, so non-parametric statistics were used. For each of the nine modeling approaches (six ANNs and three regression equations), predicted EE was averaged across epochs, separately for each activity (rest/transition and in each of the four exergames). Predicted EE from each model for each activity was compared to the criterion using a related-samples Friedman analysis of variance. In the event of a significant overall test statistic, *post hoc* differences between model predictions and the criterion were evaluated using pairwise, related-samples Wilcoxon rank sum tests.

For each epoch, squared error was calculated to compare each of the nine modeling approaches (six ANNs, three regression equations) to the criterion EE. Then, for each participant, root mean squared error (RMSE) was calculated, separately for each model. A related-samples Friedman analysis of variance test was used to compare RMSE across ANN and regression models, with *post hoc* differences between models evaluated using a pairwise, related-samples Wilcoxon rank sum test. A *p*-value <.05 was used to indicate statistical significance, and a false discovery rate correction was used to account for multiple comparisons (Glickman, Rao, & Schultz, 2014). Bland-Altman plots (Bland & Altman, 1986) were also created to evaluate bias in EE prediction across the nine models evaluated. These plots revealed an outlier for the three raw data ANNs. To determine if the outlier data affected our findings, we ran both Friedman analyses (the analysis comparing criterion EE to predicted EE for each activity and the analysis comparing RMSE among model types) twice, once with outlier data included and once with outlier data excluded. There were no changes in the statistical significance of the findings for either Friedman analysis, so data in the Results are shown with the outlier data included. Analyses were conducted using SPSS version 24.0.

Results

RMSE for EE prediction for each modeling approach is shown in Figure 1. The Friedman test statistic was statistically significant; *post hoc* analyses for RMSE revealed that the count-based, hip ANN had significantly lower RMSE than all other ANNs and regression models. More specifically, the count-based hip ANN had RMSE (mean ± standard deviation of 1.2 ± 0.3 METs) 19.2% lower than the next best model (hip regression; 1.5 ± 0.5 METs). Conversely, the raw wrist ANN (RMSE: 3.6 ± 1.7 METs) had significantly higher RMSE than all other ANNs and all regression models except the wrist regression model (RMSE: 3.2 ± 0.8 METs). For both the hip and wrist ANNs, the count-based models had significantly lower RMSE (39.6%–41.0% lower) than the raw models, although this was not the case for the combination ANNs. In comparing ANNs to the regression models, the count-based ANNs for each accelerometer placement (and combination) had significantly lower RMSE than their corresponding regression equations, and the raw ANNs had significantly higher RMSE than their corresponding regression equations.

Bland-Altman plots (Figure 2) revealed "bias toward the mean", where all nine predictive models overestimated EE when criterion-measured EE was low (i.e., during low-intensity activities) and underestimated EE when criterion-measured EE was high (i.e., during high-intensity activities). Additionally, these plots revealed narrower 95% limits of agreement for all three count-based ANNs compared to the raw ANNs for the hip, wrist, and combination ANNs. Limits of agreement for the hip, wrist, and combination regression models were wider than that of the count-based ANNs but narrower than the raw data ANNs. As indicated in the Methods, these plots revealed an outlier in the dataset for one participant, for which EE was substantially overestimated by all three raw ANNs (average overestimation of 4.6 METs for hip, 10.3 METs for wrist, and 6.0 METs for combination). This participant's data were removed and the data reanalyzed, which reduced RMSE to 1.9 METs (from 2.0 METs) for the raw hip ANN, 3.4 METs (from 3.6 METs) for the raw wrist ANN, and 1.7 METs (from 1.9 METs) for the raw combination ANN, but there was no change in

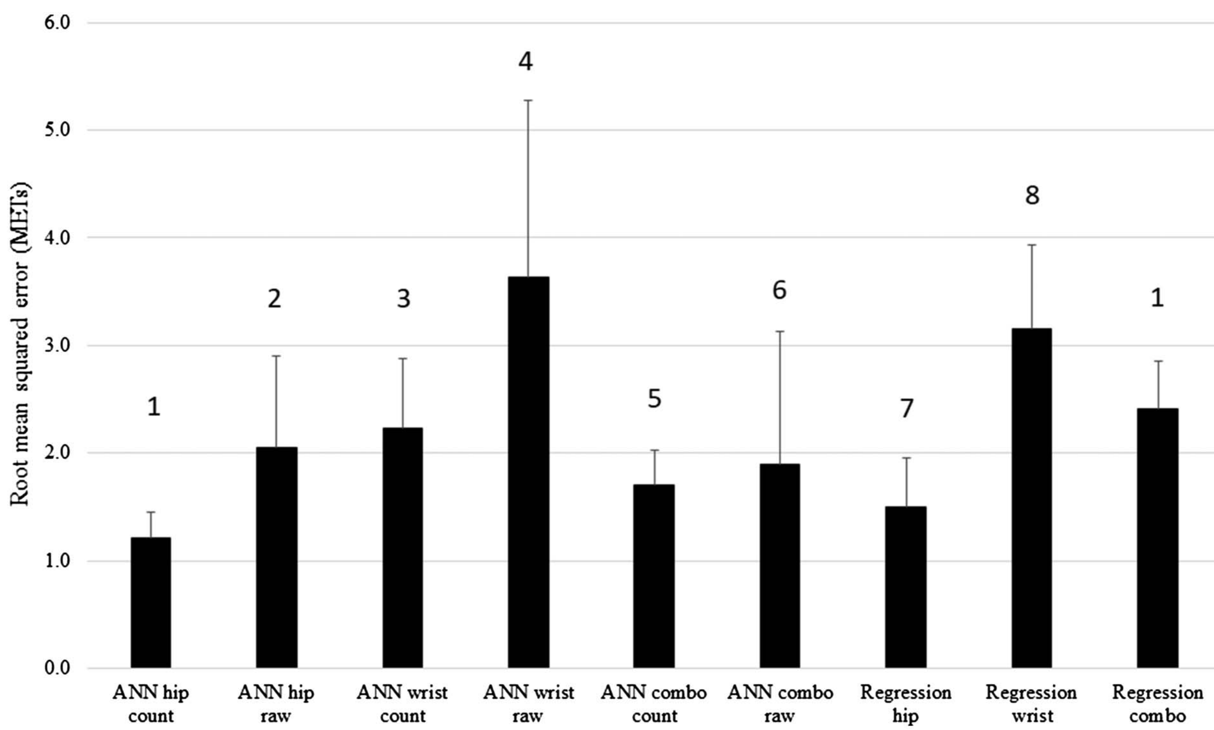

Figure 1 — Root mean squared error for energy expenditure prediction. [1]Indicates significant difference from all other models. [2]Indicates significant difference from all models except ANN wrist count and ANN combination raw models. [3]Indicates significant difference from all models except ANN hip raw and ANN combination raw models. [4]Indicates significant difference from all models except regression wrist model. [5]Indicates significant difference from all models except ANN combination raw model. [6]Indicates significant difference from ANN hip count, ANN wrist raw, regression wrist, and regression combination models. [7]Indicates significant difference from all models except ANN combination raw model. [8]Indicates significant difference from all models except ANN wrist raw model. ANN = artificial neural network, Combo = Combination of hip and wrist data, METs = Metabolic equivalents.

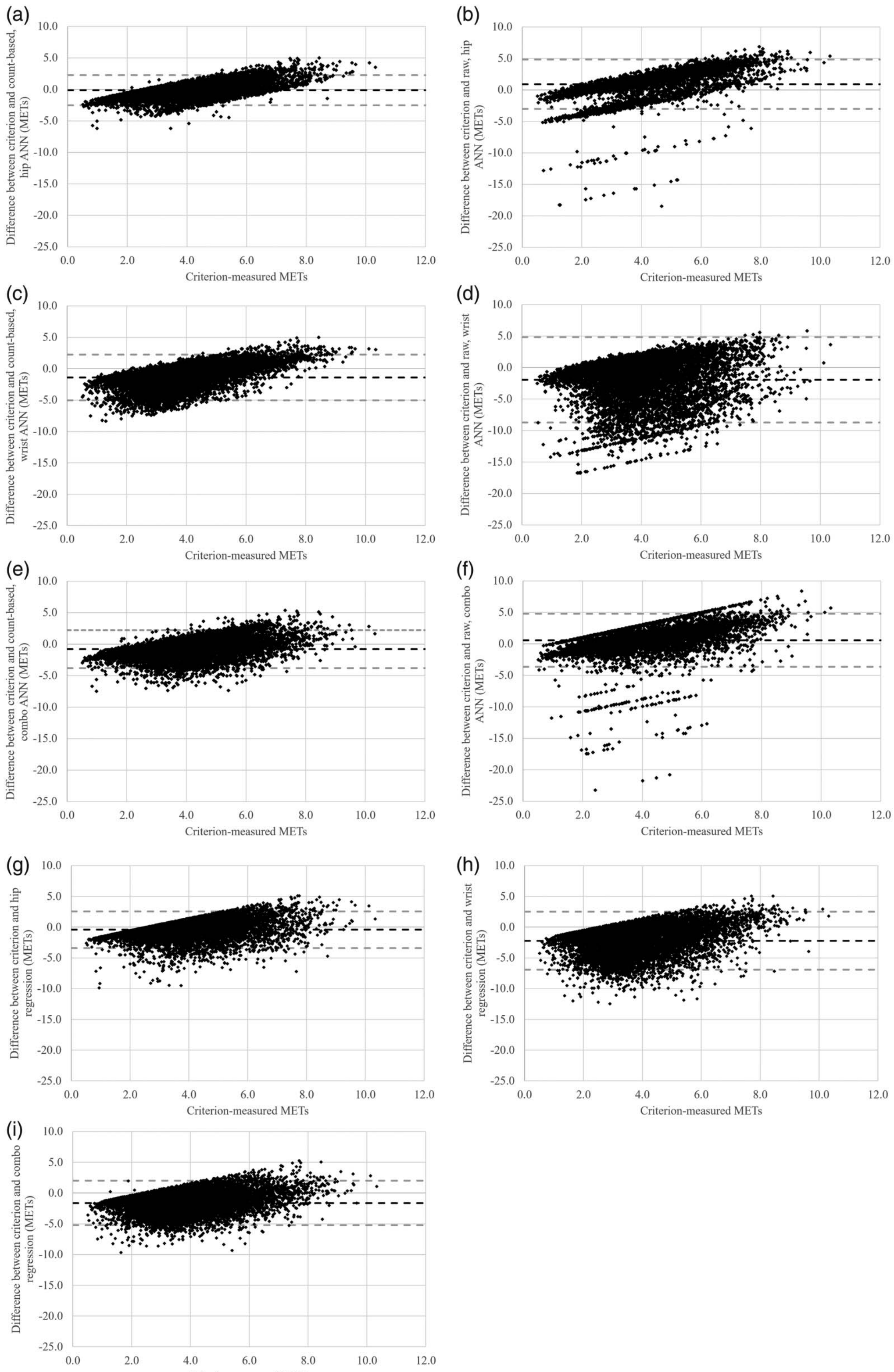

Figure 2 — Bland-Altman plots showing agreement between predicted and measured energy expenditure when cross-validating artificial neural networks and regression models. a: ANN developed from count-based, hip accelerometer data. b: ANN developed from raw, hip accelerometer data. c: ANN developed from count-based, wrist accelerometer data. d: ANN developed from raw, wrist accelerometer data. e: ANN developed from count-based, combination accelerometer data. f: ANN developed from raw, combination accelerometer data. g: Regression model developed from count-based, hip accelerometer data. h: Regression model developed from count-based, wrist accelerometer data. i: Regression model from count-based, combination accelerometer data. ANN = artificial neural network, Combo = Combination of hip and wrist data, METs = Metabolic equivalents. Points greater than 0 on the *y*-axis represent underestimation by the predictive model, and vice versa for points less than 0.

the overall findings. Similarly, removal of the outlier lowered the limits of agreement (shown as low, high) for the raw hip, wrist, and combination ANNs [(−4.3, 2.2 METs), (−4.5, 7.8 METs), and (−3.8, 2.3 METs), respectively] but did not affect overall comparisons.

Criterion-measured and accelerometer-predicted EE for each activity in cross-validation are shown in Table 2. Criterion data from the calibration dataset were comparable to that of cross-validation, with an EE of 2.8 ± 1.6 METs during rest/transitions, 3.7 ± 1.2 METs during the exergame Kinect River Rush, 4.6 ± 1.6 METs during the exergame Kinect Adventures Reflex Ridge, and 6.5 ± 2.6 METs during the treadmill test in calibration. In line with the overall RMSE analysis, the count-based hip ANN performed best; while predicted EE was significantly different from the criterion at rest and for two (of four) exergames, average EE predictions were within 0.6 METs of the criterion measure for all activities and were not different from the criterion overall. Average EE from the hip regression equation was the next best, with EE within 0.7 METs of the criterion for all activities (although all were statistically different from the criterion). Conversely, both wrist ANNs, the count-based combination ANN, the wrist regression equation, and the hip-wrist combination regression equation significantly overestimated EE overall and for all activities, with biases of 1.4 to 2.2 METs overall, 0.5 to 1.6 METs for rest/transitions, and 0.4 to 4.1 METs during the exergames.

Discussion

Our study used a semi-structured setting primarily involving exergame play to determine whether data type (count-based or raw) and/or accelerometer placement (hip, wrist, or combination) affect ANN-based EE prediction accuracy in youth and how such EE prediction models compared to count-based regression equations. Overall, an ANN machine learning model using count-based, hip accelerometer data had lower error in predicting EE compared to ANNs developed using wrist or combination data, with RMSE 29.2% lower than the next best performing ANN. Notably, count-based ANNs generally outperformed raw data ANNs. Additionally, the count-based, hip ANN had lower error than three count-based EE regression equations (RMSE 19.2% lower than the best performing regression model), which is consistent with past work showing that machine learning methods may improve EE prediction compared to simple regression models (Montoye et al., 2015; Staudenmayer et al., 2009).

We report lower error of predictive models using count-based, hip-worn accelerometer data compared to wrist-worn or combination of hip- and wrist-worn accelerometers. Comparisons of models using data from hip- versus wrist-worn accelerometers for predicting EE have had equivocal results, with studies across youth and adult samples indicating lower RMSE values from either hip (Hibbing, LaMunion, Kaplan, & Crouter, 2018; Mackintosh et al., 2016) or wrist (Crouter et al., 2015; Ellis et al., 2014; Staudenmayer et al., 2015) wear locations, or indicating that which model performed better depended on the input features (Montoye et al., 2015). Of note, these previous studies have generally reported small differences in RMSE between hip-and wrist-worn monitors (e.g., 0.1–0.2 METs), in contrast to the larger differences in RMSE in the present study (1.2–3.6 METs).

The larger RMSE values in the present study compared to previous studies may be due to our focus on exergames, which involve sporadic arm movements. Graves et al. (2008) previously found that hip-worn accelerometers could better predict EE than accelerometers placed on the upper limb during exergames, indicating that the higher accuracy of hip-worn models may be due to the types of activities included in the present study. Additionally, Hwang, Fernandez, and Lu (2018) reported poorer reliability for ActiGraph monitors worn on the wrist compared to the hip during exergames, further supporting that the poorer performance of the wrist-worn monitor may be at least partially due to the focus of the present study on exergames.

It should also be considered that which wrist the accelerometer is worn on may be important when wearing an accelerometer during exergame play. Graves et al. (2008) found that non-dominant arm movement was largely impacted by the type of exergame youth participate in and/or their skill level, while dominant arm movement was not. While we did not assess participant handedness in our samples, population estimates suggest that only ~8% of individuals are left-hand dominant (McManus, 1991), so our left-wrist accelerometer was likely the non-dominant wrist for the vast majority of the sample. Future research should therefore ascertain whether an accelerometer worn on the dominant wrist would be preferable for improving EE prediction accuracy in this setting.

Table 2 Criterion-Measured and Accelerometer-Predicted Energy Expenditure in Cross-Validation

	Overall (*N* = 34)	Rest/transition (*n* = 34)	Kinect Adventures Reflex Ridge (n = 17)	Just Dance 3 (n = 17)	Wipeout (n = 16)	Kinect Sports Boxing (n = 18)
Criterion	3.9 (1.5)	2.7 (1.1)	4.4 (1.5)	4.4 (1.5)	4.0 (1.5)	3.2 (1.0)
Hip count ANN	4.0 (1.2)	3.3 (0.7)*	4.6 (1.4)	4.4 (1.2)	3.7 (0.8)*	3.7 (0.8)*
Hip raw ANN	3.0 (1.6)*	2.7 (1.4)	3.1 (1.9)*	3.3 (1.5)*	2.9 (1.7)*	2.9 (1.4)*
Wrist count ANN	5.3 (1.6)*	4.1 (1.2)*	4.9 (1.2)*	5.3 (1.3)*	5.2 (1.4)*	6.5 (1.7)*
Wrist raw ANN	5.9 (3.7)*	3.8 (2.8)*	4.8 (3.1)*	6.2 (3.4)*	7.6 (4.5)*	5.6 (2.9)*
Combination count ANN	4.7 (1.7)	3.5 (1.0)*	4.9 (1.6)*	5.1 (1.7)*	4.4 (1.6)*	4.9 (1.6)*
Combination raw ANN	3.3 (2.0)	2.8 (1.7)	3.6 (2.5)*	3.3 (1.6)*	3.9 (2.3)	2.7 (1.1)*
Regression hip VM	4.3 (1.7)	3.2 (1.1)*	5.1 (2.2)*	4.8 (1.8)*	4.0 (1.3)	3.7 (1.0)*
Regression wrist VM	6.1 (2.3)*	4.3 (1.7)*	5.4 (1.7)*	5.8 (1.9)*	6.7 (2.9)*	7.3 (2.1)*
Regression combination VM	5.5 (2.0)*	3.8 (1.4)*	5.5 (1.8)*	5.4 (1.7)*	5.4 (2.0)*	5.4 (1.5)*

Note. Data are shown in metabolic equivalents (METs), as mean (standard deviation). Combination = Combination of hip and wrist data. ANN = Artificial neural network machine learning model. VM = regression equation developed using vector magnitude of count-based data.
*Indicates significant difference from the criterion.

Better accuracy of the hip ANN in the independent sample in the present study may also be due to less movement variability at the hip compared to the wrist among participants and among different exergames. Differences in movement variability between hip and wrists are likely also present with other non-ambulatory activities that take place in free-living settings. Variability in wrist movements during this type of activity may have also contributed to the poorer accuracy of the wrist and combination in this independent sample cross-validation compared to our previous calibration study in the same setting and the same population (Mackintosh et al., 2016), whereas the hip ANNs were affected but to a lesser degree. Despite interest in wrist-worn accelerometers for the purposes of improved compliance (Troiano et al., 2014) and/or measurement of other health-related behaviors such as sleep (van Hees et al., 2015), more work is needed to improve EE prediction accuracy of wrist-worn activity monitors. A recent meta-analysis by O'Driscoll et al. (2018) suggests that the accuracy of wrist- and arm-worn monitors for predicting EE is improved with the addition of physiological data such as heart rate, so future work should evaluate this and other additional sensing methods as a potential way to improve wrist-based EE prediction.

A second important finding is that ANNs developed independently from hip- and wrist-worn, count-based data had lower error than corresponding raw data ANNs. This may be related to count-based data being designed specifically to capture acceleration frequency/magnitude and to filter accelerations that occur outside of a certain range (ActiGraph, 2016; Brønd & Arvidsson, 2016). Indeed, the conversion of raw data into counts may reduce instances where aberrant movements unduly affect EE prediction. However, despite the superior performance of count-based ANNs compared to raw data ANNs in the present study, it is pertinent to note that counts are a manufacturer-specific metric that cannot easily be translated or compared across accelerometer brands (John & Freedson, 2012), contrary to raw data which should be similar. The proprietary nature of count generation and the non-comparability of count-based data across brands render count-based models of limited use, unless 1) ActiGraph monitors are used or 2) count data equivalent to the ActiGraph are generated from the raw data of other accelerometer brands, which is now possible due to recent work by Brønd et al. (2017). Nonetheless, the higher accuracy of count-based than raw data models found in this study is informative and may offer researchers information as to how to improve the accuracy of raw data modeling techniques. Future research could investigate filtering methods, other features such as frequency-domain features, and possible translation of raw data into orientation-independent metrics such as VM, Euclidean norm minus one, or mean amplitude deviation. Such methods may allow for the use of the meaningful aspects of raw data while also removing signal noise (Bai et al., 2016; Bakrania et al., 2016). Additionally, by making these methods open-access, the comparability of data across brands would be preserved, allowing predictive models to be used across accelerometer brands.

A final notable finding is that the ANNs developed from a combination of hip and wrist data had poorer accuracy than the hip ANNs and hip regression equations in our study. Findings comparing accuracy of single- and multiple-accelerometer prediction methods are mixed. For example, two studies by Dong et al. (Dong, Biswas, Montoye, & Pfeiffer, 2013; Dong, Montoye, Moore, Pfeiffer, & Biswas, 2013) found that a three-accelerometer system (wrist, thigh, ankle) improved percent agreement for activity type classification over any single accelerometer but did not improve EE prediction over a thigh-worn accelerometer in an adult sample. Additionally, studies examining the IDEEA monitor, a five-accelerometer system (left and right upper leg, left and right foot, sternum), generally show better EE prediction than some, but not all, single-accelerometer prediction models in adults (Dannecker, Sazonova, Melanson, Sazonov, & Browning, 2013; Lof, Henriksson, & Forsum, 2013; Ryan & Gormley, 2013). In contrast, a previous study from our group (Mackintosh et al., 2016) demonstrated no additional EE prediction accuracy in youth when combining up to eight combinations of accelerometer locations relative to either hip or wrist alone in youth. Given that multi-accelerometer systems may provide only small, if any, additional EE prediction accuracy, their utility may be limited for EE prediction given the additional burden to researchers as well as participants.

Our study has several notable strengths. Specifically, the direct comparison of two popular accelerometer placement sites and their combination as well as count-based and raw data offers important considerations for how to use accelerometers for EE prediction. Additionally, the use of an independent sample for cross-validation is a strength of the present study. Previous studies that aimed to develop machine learning models for EE prediction have often used leave-one-out, *k*-fold, or other similar holdout development/testing methods, which allow for training and testing of models to be conducted efficiently within small samples. Because training and testing is being conducted using data from the same study, there is inherent similarity in the types of activities performed, setting, available equipment, and participant recruitment (Shao, 1993). Unsurprisingly, activity type or intensity prediction models developed in youth and adult samples have yielded lower accuracy when evaluated in an independent cross-validation (Gyllensten & Bonomi, 2011; Hibbing, Ellingson, et al., 2018; Kerr et al., 2016; Sasaki et al., 2016). While our cross-validation sample had overlap in one of the four exergames used compared to the calibration sample, there were still differences in several of the activities (e.g., graded exercise test in calibration only), setting, recruitment, and equipment that make the cross-validation sample independent from the sample used to calibrate the models. However, future studies should aim to cross-validate these models in independent samples participating in a larger variety of activities (e.g., ball games, tag).

Several study limitations must also be acknowledged. The present study was a secondary analysis of data from a protocol conducted in a laboratory setting using only exergames and, in calibration, a graded exercise test, resulting in limited activity types as well as a high proportion of time spent in MVPA and low time spent sedentary. Therefore, the applicability of the developed ANNs to other activity types, less active portions of a youth's day, or for free-living EE prediction is unknown. Given the overestimation of EE during rest/transitions by most ANNs and the regression models during this study as well as the intercepts for the regression models falling close to 2.5 METs, these models are not suited to detecting time spent in sedentary behavior and are, therefore, only potentially useable for predicting MVPA. As such, these models will have poorer accuracy if used to predict EE across a full day. Second, our comparison of count-based and raw data, as well as different accelerometer placements, used only one type of machine learning model, one set of features, and one epoch length. Performance of different types of models, different feature sets, and across different epoch lengths may yield informative results and should be explored in future research. Due to the more varied movements at the wrist than the hip, it may be that more complex features of the raw accelerometer data and/or other sensor inputs such as gyroscope, barometric pressure, or heart rate may aid in EE prediction accuracy from wrist-worn accelerometers (Wang et al., 2012). Finally, our raw accelerometer data were not autocalibrated, as is recommended with raw data collection (van Hees et al., 2014), due to too short of a data collection session in our laboratory setting. Autocalibration in a similar dataset from our research team revealed calibration errors of ~2.2% which, although minor, could potentially impact the machine learning models developed from raw data, so this should be evaluated in future work.

Conclusions

In summary, our study found that a machine learning model developed from count-based, hip accelerometer data had higher EE prediction accuracy in youth during an exergame session than count-based models developed from wrist data or a combination of hip and wrist data and higher accuracy than corresponding raw data models and count-based regression equations. Although our results should be confirmed using other types of machine learning models and feature sets, as well as being expanded to include activities other than exergames, our preliminary findings suggest that a hip-worn accelerometer will provide better accuracy than wrist-worn accelerometers for EE assessment in youth for assessing MVPA during exergames. On a separate note, we recommend that transparent methods for filtering and processing raw accelerometer data be developed to improve accuracy and comparability of accelerometer-based EE prediction.

Acknowledgments

The authors would like to acknowledge David Barnett and Frazer Ashman as well as funding from Bridging the Gaps for the Swansea portion of the study data, and Wei Peng, Jourdin Barkman, Allie Diltz, and funding from the Michigan State University Department of Media and Information for the Michigan State University portion of the study data. Additionally, the authors would like to thank Jan Brønd and Daniel Arvidsson for resampling the data from our calibration dataset so that it would be comparable to the cross-validation dataset.

References

ActiGraph. (2016). *What are counts?* Retrieved from https://actigraphcorp.force.com/support/s/article/What-are-counts

Bai, J., Di, C., Xiao, L., Evenson, K.R., LaCroix, A.Z., Crainiceanu, C.M., & Buchner, D.M. (2016). An activity index for raw accelerometry data and its comparison with other activity metrics. *PLoS One, 11,* e0160644. PubMed ID: 27513333. doi:10.1371/journal.pone.0160644

Bailey, R.C., Olson, J., Pepper, S.L., Porszasz, J., Barstow, T.J., & Cooper, D.M. (1995). The level and tempo of children's physical activities: An observational study. *Medicine & Science in Sports & Exercise, 27*(7), 1033–1041. PubMed ID: 7564970. doi:10.1249/00005768-199507000-00012

Bakrania, K., Yates, T., Rowlands, A.V., Esliger, D.W., Bunnewell, S., Sanders, J., . . . Edwardson, C.L. (2016). Intensity thresholds on raw acceleration data: Euclidean norm minus one (ENMO) and mean amplitude deviation (MAD) approaches. *PLoS One, 11,* e0164045. PubMed ID: 27706241. doi:10.1371/journal.pone.0164045

Barkman, J., Pfeiffer, K.A., Diltz, A., & Peng, W. (2016). Examining energy expenditure in youth Using Xbox Kinect: Differences by player mode. *Journal of Physical Activity and Health, 13,* S41–S43. PubMed ID: 27392377. doi:10.1123/jpah.2016-0016

Bland, J.M., & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307–310. PubMed ID: 2868172

Bouten, C.V., Sauren, A.A., Verduin, M., & Janssen, J.D. (1997). Effects of placement and orientation of body-fixed accelerometers on the assessment of energy expenditure during walking. *Medical & Biological Engineering & Computing, 35*(1), 50–56. PubMed ID: 9136191. doi:10.1007/BF02510392

Brønd, J.C., Andersen, L.B., & Arvidsson, D. (2017). Generating ActiGraph counts from raw acceleration recorded by an alternative monitor. *Medicine & Science in Sports & Exercise, 49*(11), 2351–2360. doi:10.1249/MSS.0000000000001344

Brønd, J.C., & Arvidsson, D. (2016). Sampling frequency affects the processing of ActiGraph raw acceleration data to activity counts. *Journal of Applied Physiology, 120*(3), 362–369. doi:10.1152/japplphysiol.00628.2015

Clevenger, K.A., & Howe, C.A. (2015). Energy cost and enjoyment of active videogames in children and teens: Xbox 360 Kinect. *Games for Health Journal, 4,* 318–324. PubMed ID: 26182220. doi:10.1089/g4h.2014.0101

Crouter, S.E., Flynn, J.I., & Bassett, D.R., Jr. (2015). Estimating physical activity in youth using a wrist accelerometer. *Medicine & Science in Sports & Exercise, 47*(5), 944–951. PubMed ID: 25207928. doi:10.1249/MSS.0000000000000502

Crouter, S.E., Horton, M., & Bassett, D.R., Jr. (2012). Use of a two-regression model for estimating energy expenditure in children. *Medicine & Science in Sports & Exercise, 44*(6), 1177–1185. PubMed ID: 22143114. doi:10.1249/MSS.0b013e3182447825

Dannecker, K.L., Sazonova, N.A., Melanson, E.L., Sazonov, E.S., & Browning, R.C. (2013). A comparison of energy expenditure estimation of several physical activity monitors. *Medicine & Science in Sports & Exercise, 45,* 2105–2112. PubMed ID: 23669877. doi:10.1249/MSS.0b013e318299d2eb

Dong, B., Montoye, A., Moore, R., Pfeiffer, K., & Biswas, S. (2013). Energy-aware activity classification using wearable sensor networks. *Proceedings of the SPIE International Society for Optical Engineering, 2013,* 87230Y.

Dong, B., Biswas, S., Montoye, A., & Pfeiffer, K. (2013). Comparing metabolic energy expenditure estimation using wearable multi-sensor network and single accelerometer. *Conference Proceedings of the IEEE Engineering in Medicine and Biology Society, 2013,* 2866–2869.

Ellis, K., Kerr, J., Godbole, S., Lanckriet, G., Wing, D., & Marshall, S. (2014). A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. *Physiological Measurement, 35*(11), 2191–2203. doi:10.1088/0967-3334/35/11/2191

Esteban-Cornejo, I., Tejero-Gonzalez, C.M., Sallis, J.F., & Veiga, O.L. (2015). Physical activity and cognition in adolescents: A systematic review. *Journal of Science and Medicine in Sport, 18,* 534–539. PubMed ID: 25108657. doi:10.1016/j.jsams.2014.07.007

FAO/WHO/UNU Expert Consultation. (2001). *Energy requirements of adults*. Retrieved from http://www.fao.org/docrep/007/y5686e/y5686e07.htm

Freedson, P., Pober, D., & Janz, K.F. (2005). Calibration of accelerometer output for children. *Medicine & Science in Sports & Exercise, 37,* S523–S530. PubMed ID: 16294115. doi:10.1249/01.mss.0000185658.28284.ba

Glickman, M.E., Rao, S.R., & Schultz, M.R. (2014). False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology, 67*(8), 850–857. PubMed ID: 24831050. doi:10.1016/j.jclinepi.2014.03.012

Graves, L.E., Ridgers, N.D., & Stratton, G. (2008). The contribution of upper limb and total body movement to adolescents' energy expenditure whilst playing Nintendo Wii. *European Journal of Applied Physiology, 104,* 617–623. PubMed ID: 18607619. doi:10.1007/s00421-008-0813-8

Gyllensten, I.C., & Bonomi, A.G. (2011). Identifying types of physical activity with a single accelerometer: Evaluating laboratory-trained algorithms in daily life. *IEEE Transactions on Biomedical Engineering, 58,* 2656–2663. PubMed ID: 21712150. doi:10.1109/TBME.2011.2160723

Hibbing, P.R., Ellingson, L.D., Dixon, P.M., & Welk, G.J. (2018). Adapted sojourn models to estimate activity intensity in youth: A suite of tools. *Medicine & Science in Sports & Exercise, 50*(4), 846–854. PubMed ID: 29135657. doi:10.1249/MSS.0000000000001486

Hibbing, P.R., LaMunion, S.R., Kaplan, A.S., & Crouter, S.E. (2018). Estimating energy expenditure with ActiGraph GT9X inertial measurement unit. *Medicine & Science in Sports & Exercise, 50*(5), 1093–1102. PubMed ID: 29271847. doi:10.1249/MSS.0000000000001532

Hwang, J., Fernandez, A.M., & Lu, A.S. (2018). Application and validation of activity monitors' epoch lengths and placement sites for physical activity assessment in exergaming. *Journal of Clinical Medicine, 7*(9), E268. PubMed ID: 30208567. doi:10.3390/jcm7090268

John, D., & Freedson, P. (2012). ActiGraph and Actical physical activity monitors: A peek under the hood. *Medicine & Science in Sports & Exercise, 44,* S86–S89. PubMed ID: 22157779. doi:10.1249/MSS.0b013e3182399f5e

Kerr, J., Patterson, R.E., 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), 951–957. PubMed ID: 27089222. doi:10.1249/MSS.0000000000000841

Lof, 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 Sciences, 2,* e31. doi:10.1017/jns.2013.18

Lyden, K., Keadle, S.K., Staudenmayer, J., & Freedson, P.S. (2014). A method to estimate free-living active and sedentary behavior from an accelerometer. *Medicine & Science in Sports & Exercise, 46,* 386–397. PubMed ID: 23860415. doi:10.1249/MSS.0b013e3182a42a2d

Lyons, R. (2013). *Understanding digital signal processing* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Mackintosh, K.A., Montoye, A.H.K., Pfeiffer, K.A., & McNarry, M. (2016). Investigating optimal accelerometer placement for energy expenditure prediction in children using a machine learning approach. *Physiological Measurement, 37,* 1728–1740. PubMed ID: 27653339. doi:10.1088/0967-3334/37/10/1728

McClain, J.J., Abraham, T.L., Brusseau, T.A., Jr, & Tudor-Locke, C. (2008). Epoch length and accelerometer outputs in children: comparison to direct observation. *Medicine & Science in Sports & Exercise, 40*(12), 2080–2087. doi:10.1249/MSS.0b013e3181824d98

McManus, C. (1991). The inheritance of left-handedness. In: G.R. Bock & J. Marsh (Eds.), *Ciba Foundation Symposium: Biological Asymmetry and Handedness*. West Sussex, England: John Wiley & Sons Ltd.

McMurray, R.G., Butte, N.F., Crouter, S.E., Trost, S.G., Pfeiffer, K.A., Bassett, D.R., … Fulton, J.E. (2015). Exploring metrics to express energy expenditure of physical activity in youth. *PLoS One, 10,* e0130869. PubMed ID: 26102204. doi:10.1371/journal.pone.0130869

Montoye, A.H.K., Begum, M., Henning, Z., & Pfeiffer, K.A. (2017). Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data. *Physiological Measurement, 38,* 343–357. PubMed ID: 28107205. doi:10.1088/1361-6579/38/2/343

Montoye, A.H.K., Moore, R.W., Bowles, H.R., Korycinski, R., & Pfeiffer, K.A. (2018). Reporting accelerometer methods in physical activity intervention studies: a systematic review and recommendations for authors. *British Journal of Sports Medicine, 52*(23), 1507–1516. doi:10.1136/bjsports-2015-095947

Montoye, A.H.K., Mudd, L.M., Biswas, S., & Pfeiffer, K.A. (2015). Energy expenditure prediction using raw accelerometer data in simulated free living. *Medicine & Science in Sports & Exercise, 47,* 1735–1746. PubMed ID: 25494392. doi:10.1249/MSS.0000000000000597

Montoye, A.H.K., Pivarnik, J.M., Mudd, L.M., Biswas, S., & Pfeiffer, K.A. (2016). Wrist-independent energy expenditure prediction models from raw accelerometer data. *Physiological Measurement, 37,* 1770–1784. PubMed ID: 27653642. doi:10.1088/0967-3334/37/10/1770

O'Driscoll, R., Turicchi, J., Beaulieu, K., Scott, S., Matu, J., Deighton, K., … 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*. Advance online publication. doi:10.1136/bjsports-2018-099643

Preece, S.J., Goulermas, J.Y., Kenney, L.P., Howard, D., Meijer, K., & Crompton, R. (2009). Activity identification using body-mounted sensors: A review of classification techniques. *Physiological Measurement, 30*(4), R1–R33. PubMed ID: 19342767 doi:10.1088/0967-3334/30/4/R01

Ripley, B., & Venables, W. (2016). *Package 'nnet': Feed-forward neural networks and multinomial log-linear models*. Retrieved from https://cran.r-project.org/web/packages/nnet/nnet.pdf

Rosdahl, H., Gullstrand, L., Salier-Eriksson, J., Johansson, P., & Schantz, P. (2010). Evaluation of the Oxycon Mobile metabolic system against the Douglas bag method. *European Journal of Applied Physiology, 109*(2), 159–171. PubMed ID: 20043228. doi:10.1007/s00421-009-1326-9

Rosenberg, M., Lay, B., Lee, M., Derbyshire, A., Kur, J., Ferguson, R., … Braham, R. (2013). New-generation active videogaming maintains energy expenditure in children across repeated bouts. *Games for Health Journal, 2,* 274–279. PubMed ID: 26196928. doi:10.1089/g4h.2013.0037

Ryan, J., & Gormley, J. (2013). An evaluation of energy expenditure estimation by three activity monitors. *European Journal of Sports Science, 13,* 681–688. doi:10.1080/17461391.2013.776639

Sasaki, J.E., John, D., & Freedson, P.S. (2011). Validation and comparison of ActiGraph activity monitors. *Journal of Science and Medicine in Sport, 14*(5), 411–416. PubMed ID: 21616714. doi:10.1016/j.jsams.2011.04.003

Sasaki, J.E., Hickey, A.M., Staudenmayer, J.W., John, D., Kent, J.A., & Freedson, P.S. (2016). Performance of activity classification algorithms in free-living older adults. *Medicine & Science in Sports & Exercise, 48,* 941–950. PubMed ID: 26673129. doi:10.1249/MSS.0000000000000844

Schofield, W.N. (1985). Predicting basal metabolic rate, new standards and review of previous work. *Human Nutrition and Clinical Nutrition, 39*(Suppl. 1), 5–41.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association, 88,* 486–494. doi:10.1080/01621459.1993.10476299

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,* 396–403. PubMed ID: 26112238. doi:10.1152/japplphysiol.00026.2015

Staudenmayer, J., Pober, D., Crouter, S., Bassett, D., & Freedson, P. (2009). An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. *Journal of Applied Physiology, 107,* 1300–1307. PubMed ID: 19644028. doi:10.1152/japplphysiol.00465.2009

Swartz, A.M., Strath, S.J., Bassett, D.R., Jr., O'Brien, W.L., King, G.A., & Ainsworth, B.E. (2000). Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. *Medicine & Science in Sports & Exercise, 32,* S450–S456. PubMed ID: 10993414. doi:10.1097/00005768-200009001-00003

Troiano, R.P., Berrigan, D., Dodd, K.W., Masse, L.C., Tilert, T., & McDowell, M. (2008). Physical activity in the United States measured by accelerometer. *Medicine & Science in Sports & Exercise, 40,* 181–188. PubMed ID: 18091006. doi:10.1249/mss.0b013e31815a51b3

Troiano, R.P., McClain, J.J., Brychta, R.J., & Chen, K.Y. (2014). Evolution of accelerometer methods for physical activity research. *British Journal of Sports Medicine, 48,* 1019–1023. PubMed ID: 24782483. doi:10.1136/bjsports-2014-093546

Trost, S.G., Wong, W.K., Pfeiffer, K.A., & Zheng, Y. (2012). Artificial neural networks to predict activity type and energy expenditure in youth. *Medicine & Science in Sports & Exercise, 44,* 1801–1809. PubMed ID: 22525766. doi:10.1249/MSS.0b013e318258ac11

US Department of Health and Human Services. (2018). *Physical Activity Guidelines Advisory Committee: 2018 Physical Activity Guidelines for Americans* (2nd ed.). Rockville, MD: Office of Disease Prevention and Health Promotion. Retrieved from www.health.gov/paguidelines

van Hees, V.T., Fang, Z., Langford, J., Assah, F., Mohammad, A., da Silva, I.C., … Brage, S. (2014). Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents. *Journal of Applied Physiology, 117*(7), 738–744. PubMed ID: 25103964. doi:10.1152/japplphysiol.00421.2014

van Hees, V.T., Gorzelniak, L., Dean Leon, E.C., Eder, M., Pias, M., Taherian, S., … Brage, S. (2013). Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PLoS One, 8*(4), e61691. PubMed ID: 23626718. doi:10.1371/journal.pone.0061691

van Hees, V.T., Sabia, S., Anderson, K.N., Denton, S.J., Oliver, J., Catt, M., … Singh-Manoux, A. (2015). A novel, open access method to assess sleep duration using a wrist-worn accelerometer. *PLoS One, 10*(11), e0142533. PubMed ID: 26569414. doi:10.1371/journal.pone.0142533

Wang, J., Redmond, S.J., Voleno, M., Narayanan, M.R., Wang, N., Cerutti, S., & Lovell, N.H. (2012). Energy expenditure estimation during normal ambulation using triaxial accelerometry and barometric pressure. *Physiological Measurement, 33*(11), 1811–1830. PubMed ID: 23110944. doi:10.1088/0967-3334/33/11/1811