Physiological Measurement

IPEM
Institute of Physics and Engineering in Medicine

PAPER

Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data

IOP Publishing | Institute of Physics and Engineering in Medicine Physiological Measurement

Physiol. Meas. **38** (2017) 343–357 doi:10.1088/1361-6579/38/2/343

Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data

Alexander H K Montoye[1,2,5,6], Munni Begum[3], Zachary Henning[3] and Karin A Pfeiffer[4]

[1] Department of Integrative Physiology and Health Science, Alma College, 614 W. Superior Alma, MI 48801, USA
[2] Clinical Exercise Physiology Program, Ball State University, 2000 W. University Ave. Muncie, IN 47306, USA
[3] Department of Mathematical Sciences, Ball State University, 2000 W. University Ave. Muncie, IN 47306, USA
[4] Department of Kinesiology, Michigan State University, 308 W. Circle Dr. East Lansing, MI 48824, USA

E-mail: montoyeah@alma.edu

Received 15 July 2016, revised 6 October 2016
Accepted for publication 1 November 2016
Published 20 January 2017

Abstract

This study had three purposes, all related to evaluating energy expenditure (EE) prediction accuracy from body-worn accelerometers: (1) compare linear regression to linear mixed models, (2) compare linear models to artificial neural network models, and (3) compare accuracy of accelerometers placed on the hip, thigh, and wrists. Forty individuals performed 13 activities in a 90 min semi-structured, laboratory-based protocol. Participants wore accelerometers on the right hip, right thigh, and both wrists and a portable metabolic analyzer (EE criterion). Four EE prediction models were developed for each accelerometer: linear regression, linear mixed, and two ANN models. EE prediction accuracy was assessed using correlations, root mean square error (RMSE), and bias and was compared across models and accelerometers using repeated-measures analysis of variance. For all accelerometer placements, there were no significant differences for correlations or RMSE between linear regression and linear mixed models (correlations: $r = 0.71$–0.88, RMSE: 1.11–1.61 METs; $p > 0.05$). For the thigh-worn accelerometer, there were

[5] Author to whom any correspondence should be addressed.
614 W. Superior, Alma, MI 48801.
[6] Data collection was carried out in the Department of Kinesiology at Michigan State University, 308 W. Circle Dr., East Lansing, MI 48824. Data processing and analysis were conducted in the Department of Mathematical Sciences at Ball State University, 2000 W. University Ave., Muncie, IN 47306.

no differences in correlations or RMSE between linear and ANN models (ANN—correlations: $r = 0.89$, RMSE: 1.07–1.08 METs. Linear models—correlations: $r = 0.88$, RMSE: 1.10–1.11 METs; $p > 0.05$). Conversely, one ANN had higher correlations and lower RMSE than both linear models for the hip (ANN—correlation: $r = 0.88$, RMSE: 1.12 METs. Linear models—correlations: $r = 0.86$, RMSE: 1.18–1.19 METs; $p < 0.05$), and both ANNs had higher correlations and lower RMSE than both linear models for the wrist-worn accelerometers (ANN—correlations: $r = 0.82$–0.84, RMSE: 1.26–1.32 METs. Linear models—correlations: $r = 0.71$–0.73, RMSE: 1.55–1.61 METs; $p < 0.01$). For studies using wrist-worn accelerometers, machine learning models offer a significant improvement in EE prediction accuracy over linear models. Conversely, linear models showed similar EE prediction accuracy to machine learning models for hip- and thigh-worn accelerometers and may be viable alternative modeling techniques for EE prediction for hip- or thigh-worn accelerometers.

Keywords: linear mixed model, artificial neural network, physical activity, machine learning, ActiGraph, GENEActiv

(Some figures may appear in colour only in the online journal)

Introduction

Body-worn accelerometers have emerged as a preferred method for measuring physical activity (PA) due to their objective data capture and ability to assess intensity of movement from acceleration data. Conventional accelerometer use includes placement of the accelerometer on the hip and translating the raw acceleration data collected into preprocessed 'activity counts', which are typically brand-specific and quantify the magnitude of acceleration captured over a given time interval (e.g. 60 s) (John and Freedson 2012, Montoye *et al* 2016b). In order to translate activity counts into a meaningful outcome variable, activity counts are used as an independent variable in regression equations (most often linear regression) developed to predict energy expenditure (EE); these equations have been shown to work well for ambulatory activities (i.e. walking and running) but have decreased accuracy for sedentary and other non-ambulatory activities (Freedson *et al* 1998, Hendelman *et al* 2000, Swartz *et al* 2000). While not yet widely adopted for use, machine learning modeling of raw accelerometer data has emerged as a method with strong potential for improving estimates of PA from accelerometers. For example, through use of machine learning methods, researchers have been able to assess the EE of non-ambulatory tasks with better accuracy than count-based regression and characterize specific activity types (Staudenmayer *et al* 2009, Dong *et al* 2013, Lyden *et al* 2014, Montoye *et al* 2015, Montoye 2016a). However, a number of unanswered questions remain regarding the potential utility of machine learning.

First, despite the potential of machine learning for improving EE measurement, the majority of past research has used count-based linear regression models as a comparison, with only a few studies developing regression models from raw data. A study by Esliger *et al* established cut-points for raw data collected from wrist- and hip-worn GENEA accelerometers across a variety of activities and found them to have similar accuracy to count-based regression methods developed for hip-worn accelerometers (Esliger *et al* 2011). Similarly, two studies by Vathsangam *et al* demonstrated high accuracy for EE prediction from raw data from a hip-worn accelerometer during treadmill walking (Vathsangam *et al* 2010, 2011); however, past work has demonstrated poor generalizability of models developed from treadmill-based protocols when applied to non-ambulatory activities (Strath *et al* 2003, Ham *et al* 2007). Therefore, while the potential for developing linear regression models from raw data has been established, it is unknown how the accuracy of linear regression and machine learning models will compare if the same data are used as model inputs. If linear regression models developed for analyzing raw accelerometer data can achieve similar accuracy to machine learning models, the linear regression models may be preferable for use since they are conceptually simpler and are easier to implement and interpret than machine learning models.

Additionally, past research examining different types of machine learning models have seen little difference in accuracy among different types used to assess PA from accelerometer data. Studies by Cleland *et al*, Parkka *et al*, and Dong *et al* examined activity type classification accuracy by a variety of different machine learning model types (artificial neural network, naïve Bayes, support vector machine, nearest neighbor, and decision trees) for several accelerometer placements and activity lists. All three studies found similar classification accuracies (94–98%, 82–86%, and 93–97%, respectively) regardless of the type of machine learning model used, with no model type consistently outperforming others (Dong *et al* 2013, Parkka *et al* 2006, Cleland 2013). Since choice of machine learning model seems to have little influence on PA measurement accuracy, other avenues must be explored in order to further improve PA measurement using machine learning methods.

One such method researchers have used is incorporating more and different types of features extracted from the raw accelerometer data for prediction. Generally, a greater number of features and a combination of time- and frequency-domain features have been shown to yield small improvements in accuracy for activity recognition and EE estimation (Mannini and Sabatini 2010, Montoye *et al* 2015, Staudenmayer *et al* 2015). However, a greater number of features increases the risk of over-fitting models to the dataset in which they were validated, which would lower the model accuracy when cross-validated on an independent sample. Current practices for development and validation of machine learning models for assessing PA include extracting features from the raw accelerometer data in small time windows (e.g. 30 s) and considering data extracted from all study participants equal when developing models. However, many acceleration signals and data windows are collected from each participant involved in a validation study, and the interrelatedness of data collected within each participant is much greater than that between participants. It is likely that considering the interrelatedness of data collected within participants will yield valuable additional information to improve PA prediction accuracy at both the individual and group level over models not incorporating this relatedness. This is not yet possible with machine learning models but can be tested with linear mixed modeling techniques. Therefore, the purpose of this study was threefold. The first purpose was to compare accuracy of EE prediction from simple linear regression models to linear mixed models to determine if relatedness of within-participant data can be incorporated to improve EE prediction accuracy. Second, linear models (linear regression and linear mixed models) were compared to machine learning models for EE prediction from raw accelerometer data to determine if linear models are a suitable alternative to machine learning for analyzing raw accelerometer data. Finally, accuracy of accelerometers worn on different body placements (hip, thigh, left wrist, right wrist) were compared to determine if accelerometer placement affected EE prediction accuracy.

Methods

Participants

Forty-four adults (22 male, 22 female) were recruited for this study. Inclusion criteria included being able to perform moderate- and vigorous-intensity PA safely, no significant orthopedic limitations, and falling in the age range of 18–44 years. This study was approved by the Michigan State University Institutional Review Board prior to participant recruitment and Ball State University prior to data analysis.

Equipment

Four accelerometers, including two brands, were used in this study. Participants wore two ActiGraph GT3X+ (ActiGraph LLC, Pensacola, FL) accelerometers, one on the right hip (anterior axillary line) and the other on the right thigh (anterior midline, 1/3 of the distance between the patella and inguinal crease). Both ActiGraph accelerometers were set to record raw, triaxial data at 40 Hz, which were reintegrated to 20 Hz after download. Participants also wore two GENEActiv (Activinsights Ltd, Kimbolton, Cambridgeshire, UK) accelerometers, one on the dorsal side of each wrist. The GENEA accelerometers were set to record acceleration data at a rate of 20 Hz.

Participants were also fitted with an Oxycon Mobile (Cardinal Health, Yorba Linda, CA) portable metabolic analyzer. The Oxycon is lightweight (950 g) and was secured to participants' backs via a shoulder harness. A breathing mask was connected to the Oxycon, allowing for collection and analysis of breath-by-breath expired gases and determination of oxygen consumption. The Oxycon has previously been shown to provide accurate measures of oxygen consumption for a range of exercise intensities and was used as the criterion measure for EE in this study (Rosdahl *et al* 2010, Akkermans *et al* 2012). Oxygen consumption data from the Oxycon were expressed in $ml \cdot kg^{-1} \cdot min^{-1}$ and converted to METs for analysis. The acceleration data for the accelerometers were time stamped and stored within the monitors and later were downloaded to a computer for analysis. The accelerometers and Oxycon were initialized at the beginning of each visit and synchronized to an external clock.

Procedure

Details of the study were discussed with each participant upon arrival to the laboratory. Written informed consent was obtained, and a PA readiness questionnaire was administered to ensure that the participant was healthy and had no contraindications to engaging in moderate- to vigorous-intensity PA. Height and weight were then measured by trained research assistants according to standardized methods (Malina 1995). Age and handedness were self-reported.

After being fitted with the four accelerometers and the Oxycon, participants engaged in a 90 min, semi-structured activity protocol, the details of which are described elsewhere (Montoye *et al* 2015, 2016c). Briefly, participants performed 13 activities of varying intensities and types, for 3–10 min each. Activities included lying down, reading a magazine/book, using a computer, standing quietly, folding laundry, sweeping confetti on the floor, walking at a leisurely pace, walking at a brisk pace, jogging, climbing and descending stairs, stationary cycling, biceps curls, and body-weight squats. All activities were self-paced. The walking and jogging activities were performed in a hallway, stationary cycling could be performed at any workload between 50–100 W, and biceps curls were conducted using a 1.4 kg weight in each hand. Additionally, participants could choose the order and exact timing of activities and could perform the same activity more than once if desired. An activity list was written on a whiteboard in the lab, and a researcher checked off activities as participants completed them. A trained research assistant observed and recorded each activity on a handheld computer while it was being performed and periodically updated participants on which activities they still needed to complete.

Data reduction and modeling

Time-domain features were extracted in 30 s, non-overlapping windows for each of the four accelerometers. The features included mean, standard deviation, minimum, maximum, covariance of adjacent windows of data, and the 10th, 25th, 50th, 75th, and 90th percentiles of the raw acceleration signal in each axis (10 features $\times$ 3 axes = 30 features), many of which have been commonly used in past studies employing machine learning models for EE prediction (Staudenmayer *et al* 2009, Trost *et al* 2012, Lyden *et al* 2014, Montoye *et al* 2015, 2016c). In order to reduce complexity of the models created, no frequency-domain features were extracted, nor was filtering of the raw accelerometer data conducted pre- or post-feature extraction. Predictive models were created separately for each of the four accelerometers and with four model types, resulting in 16 models developed and tested (4 model types $\times$ 4 accelerometers).

Breath-by-breath Oxycon data from the protocol were reintegrated into 30 s windows for time-matching to features extracted from the accelerometer data. Relative oxygen consumption ($\text{ml}\cdot\text{kg}^{-1}\cdot\text{min}^{-1}$) was converted into METs, by dividing by 3.5. All data from activities and transitions between activities were included in model development and testing. Therefore, both steady-state and non-steady-state data were used in development and testing of the models.

Four predictive models were built based on three statistical modeling approaches. These are (i) linear regression models, (ii) linear mixed models, and (iii) artificial neural networks.

Linear regression models (LM) were the simplest model tested. The LM were created for the response (outcome) variable METs measurement for i th participant, at jth feature (input variables), and for the k th accelerometer, denoted by y_{ijk}, can be seen in equation (1)) below: where β_{0k} is the intercept coefficient, β_{jk}, $j = 1, 2, ..30$ are the regression coefficients, x_{0k}, $(k = 1, 2, 3, 4)$ is a vector of ones with dimension same as rows in the design matrix, x_{jk} is the jth ($j = 1, 2, \ldots 30$) feature for the k th ($k = 1, 2, 3, 4$) accelerometer, and finally ϵ_{ijk} is the random noise for each observation which is normally distributed with zero mean and variance σ^2.

$$y_{ijk} = \sum_{j=0}^{30} \beta_{jk} x_{jk} + \epsilon_{ijk};\ k = 1, 2, 3, 4, \tag{1}$$

Linear mixed models (LMM) were included for analysis as they incorporate the interrelatedness of data collected from the same participant. LMM were created with the response variable METs measurement for ith participant, at jth feature and for the k th accelerometer, denoted by y_{ijk}, can be seen in equation (2) below. To incorporate interrelatedness of the data collected over time from each participant, LMM considers the intercept coefficient β_{0ik} as random, meaning that each participant has his or her own trajectory of regression. β_{0ik} is assumed to be normally distributed with mean β_{0k} and some variance σ^2_{0k}. The interrelatedness of the data from each participant is explained using estimated σ^2_{0k}.

$$y_{ijk=}\ \beta_{0ik} + \sum_{j=1}^{30} \beta_{jk} x_{jk};\ k = 1, 2, 3, 4. \tag{2}$$

Artificial neural networks (ANNs) were the third model type tested in this study. ANNs approximate the functional relationship between the response (METs) and a set of covariates (input features) without imposing a specific type of relationship (such as linear in the case of LM or LMM) between covariates and the response. The formula for an ANN with one hidden layer with J neurons ($J = 5$ in our study) can be seen in equation (3) below, where $o(\vec{x})$ is the output or response function, $\vec{x}$ is a set of covariates (30 features in our study) and w 's can be interpreted as weights similar to regression coefficients as in LM or LMM (Gunther and Fritch 2010).

$$o(\vec{x}) = f\left(w_0 + \sum_{l=1}^{5}\ w_l\ .f\left(w_{0j} + \vec{w_l}\vec{x}\right)\right) \tag{3}$$

Using these three modeling techniques, four models were developed for EE prediction from the accelerometers. The LM models tested in this study were fitted using the '*lm*' command in R computational environment (R-project, Vienna, Austria). In order to determine if utilizing within-person relatedness of data could improve predictive accuracy, LMM models were fitted using the '*lme*' command in R. Finally, ANNs were developed using two different packages in R: '*nnet*' and '*neuralnet*'. For the ANNs, we chose to use five hidden units (neurons) in the hidden layer based on the number of activities in the study and the number of features being used (Preece *et al* 2009). For the *nnet* ANNs, skip-layer connections were not allowed, and a Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization algorithm was used, as is the standard in the nnet package in R. For the *neuralnet* ANNs, the backpropagation algorithm and three versions of resilient backpropagation were implemented to fit the model (Gunther and Fritch 2010). Each of the four models was created and tested using a leave-one-out cross-validation, using the 30 features described earlier as inputs into the models. Therefore, all models were developed to predict the same response variable (METs) with the same features used as inputs. Example data and code for using the developed models can be found at the following link: https://github.com/zshenning/Accel_Model_Testing.

Statistical analyses

Pearson correlations, root mean square error (RMSE), and bias were calculated to test the accuracy of each of the four models, LM, LMM, ANN (using *nnet*) and ANN (using *neuralnet*) for predicting EE. These statistics were calculated separately for each of the models and accelerometers. Comparisons of model accuracy were assessed in three ways: (1) comparison of the linear regression model to the linear mixed model within each accelerometer placement; (2) comparison of the linear models to the ANN models within each accelerometer placement; and (3) comparison of the model types for all accelerometer placements to the traditional, Freedson MET prediction equation for the hip-worn accelerometer. Differences among correlations, RMSE, and bias were assessed using repeated-measures analysis of variance (RMANOVA). When the RMANOVA revealed statistically significant differences for any of the analyses, post hoc dependent t-tests and a false discovery rate correction were conducted to determine differences among monitor placements or model types (Glickman *et al* 2014). Statistical significance was set as an Alpha level of $P < 0.05$. Statistical analyses were performed using R and SPSS version 23.0 (IBM Corporation, Armonk, NY).

Table 1. Demographic characteristics of sample.

	Males ($n = 19$)	Females ($n = 21$)	All ($n = 40$)
Age (years)	23.7 (5.0)	20.5 (2.7)	22.0 (4.2)
Weight (kg)	84.5 (13.1)	60.4 (8.8)	71.9 (16.3)
Height (cm)	179.1 (7.7)	163.7 (5.9)	171.0 (10.3)
BMI (kg m^{-2})	26.3 (3.4)	22.5 (2.6)	24.3 (3.5)

Note: Data are displayed as mean (standard deviation).

Results

Of the 44 participants tested in the study, three had significant data loss from the Oxycon metabolic analyzer, resulting in premature termination of the protocol; additionally, an accelerometer malfunction occurred for one participant. Data from these four participants were excluded from analysis, resulting in 40 participants' data available for analysis. Demographic characteristics of the individuals included in the analysis are shown in table 1.

Boxplots for correlations, RMSE, and bias for each accelerometer placement and modeling type are shown in figures 1–3. Correlations ranged from 0.82–0.89 and RMSE ranged from 1.07–1.31 METs for the four accelerometers with the ANN models (*nnet* and *neuralnet*), whereas correlations ranged from 0.71–0.88 and RMSE ranged from 1.11–1.61 METs for the four accelerometers with the linear models (LM and LMM).

RMANOVA analyses comparing correlations among model types were statistically significant for the accelerometers worn on the hip ($F = 5.805$, $p = 0.002$), left wrist ($F = 56.582$, $p < 0.001$), and right wrist ($F = 41.041$, $p < 0.001$) but not the thigh ($F = 2.034$, $p = 0.126$). Similarly, RMANOVA analyses comparing RMSE among model types were statistically significant for the accelerometers worn on the hip ($F = 5.233$, $p = 0.004$), left wrist ($F = 47.770$, $p < 0.001$), and right wrist ($F = 29.600$, $p < 0.001$) but not the thigh ($F = 2.787$, $p = 0.054$). Post hoc pairwise comparisons of the LM models and LMM models within each accelerometer placement revealed no differences in correlations or RMSE for any of the accelerometers. Similarly, the *nnet* and *neuralnet* ANN models were not significantly different within any accelerometer placement, which was expected. Both *nnet* and *neuralnet* implement the principle of ANNs to approximate the functional relationship between the response (METs) and a set of covariates as discussed under the *Data Reduction and Modeling* section.

When comparing linear to ANN models, both ANNs had significantly higher correlations and lower RMSE than the linear models for the left wrist- and right wrist-worn accelerometers. For the hip-worn accelerometer, the *nnet* ANN had higher correlations and lower RMSE than the linear models, but the *neuralnet* ANN did not. It should also be noted that the magnitude of differences in accuracy between the linear models and ANNs was much greater for the wrist-worn accelerometers than for the hip-worn accelerometer. There were no differences in correlations or RMSE among the four models for the thigh-worn accelerometer. None of the models developed in this study had overall bias for prediction of EE.

For the LM models, the hip- and thigh-worn accelerometers had significantly higher correlations and lower RMSE than the wrist-worn accelerometers; additionally, the thigh-worn accelerometer had a significantly higher correlation than the hip-worn accelerometer and but was not significantly different for RMSE. For the LMM models, the hip- and thigh-worn accelerometers had significantly higher correlations and lower RMSE than the wrist-worn accelerometers, and the thigh-worn accelerometer had a significantly higher correlation and lower RMSE than the hip-worn accelerometer. For both the LM and LMM models, there were no statistically significant differences in correlations or RMSE between the left- and right-wrist worn accelerometers.

The results from the ANN models developed using the *nnet* package show that the thigh-worn accelerometer has significantly higher correlations and lower RMSE than both wrist-worn accelerometers but was not statistically different than the hip-worn accelerometer ($P = 0.088$ for trend toward lower RMSE for thigh-worn accelerometer). The hip-worn accelerometer had a significantly higher correlation than the right wrist-worn accelerometer but was not significantly different than the left wrist-worn accelerometer ($P = 0.063$ for trend toward higher correlation for hip-worn accelerometer); moreover, the hip-worn accelerometer had significantly lower RMSE than both wrist-worn accelerometers. Correlations and RMSE were not significantly different between the wrist-worn accelerometers.

Lastly, the results from the ANN models developed using the *neuralnet* package showed that the hip- and thigh-worn accelerometers had significantly higher correlations and lower RMSE than the wrist-worn accelerometers, although the magnitude of differences were lower than for the linear models. Additionally, there were no significant differences in correlations or RMSE between the hip- and thigh-worn accelerometers, nor were correlations or RMSE different between the left and right wrist-worn accelerometers.

Data from the Freedson MET prediction equation (Freedson *et al* 1998) from this study were reported in our previous work with 39 of the 40 participants used in the current analysis (Montoye *et al* 2015). With a correlation of 0.80 ± 0.05, an RMSE of 1.50 ± 0.27 METs, and a bias of -109.69 ± 70.50 METs, the Freedson equation had lower correlations and higher RMSE than the ANN models for all accelerometers and the linear models (both LM and LMM) for the hip- and thigh-worn accelerometers. For the wrist-worn accelerometers, the *nnet* ANN models had significantly higher correlations and lower RMSE than the Freedson MET prediction equation. The *neuralnet* ANN models for both wrist-worn accelerometers also had significantly lower RMSE, but the correlation for only the left wrist-worn accelerometer was significantly higher than the Freedson MET prediction equation. Conversely, the linear models for both wrist-worn accelerometers had significantly lower correlations with measured EE but no difference in RMSE compared to the Freedson MET prediction equation. The Freedson MET prediction equation was the only model tested with significant ($p < 0.001$) systematic bias (underestimation) across the study participants.

Discussion

This study's purpose was to evaluate different modeling types for prediction of EE from raw accelerometer data collected by accelerometers worn on several body locations. One outcome of our study was to evaluate accuracy of LM models compared to LMM models; our hypothesis was that LMM models, which take into account interrelatedness of data collected from a participant, would allow for improved EE prediction accuracy. Our hypothesis was not supported, with no differences seen between the LM and LMM models for any of the four accelerometer placements. Nonetheless, LMMs provide an efficient approach to estimating regression coefficients and enhance the precision of the predictive model. Modeling approaches using LMM take into account the interrelatedness of the response measurements collected from each participant over time. ANNs or other machine learning approaches could be extended to include the interrelatedness among the response measures collected over time. This possibility was outside the scope of this study but should be evaluated with future research.

Another primary study outcome was to compare linear models to ANN models for EE prediction. Traditionally, LM models developed for activity count data from accelerometers have been consistently outperformed by newer, machine learning approaches for modeling accelerometer data. Therefore, we expected the ANN models developed in this study have higher accuracy than LM and LMM models developed for EE prediction. However, we expected differences between linear and ANN model to be smaller than seen in previous studies since the LM and LMM models developed in our study used the same features for developing all models and three axes of data, whereas traditional LM equations have been developed for activity count-based data, most often for only the vertical acceleration axis (Freedson *et al* 1998, Swartz *et al* 2000, Crouter *et al* 2006). For the wrist-worn accelerometers, RMSE was 21–23% higher and correlations were 13–14% lower for the LM and LMM models than for the ANN models. Trends for the hip-worn accelerometer were similar, albeit differences were smaller, with RMSE 3–7% lower and correlations 2–3% higher for the results obtained from the *nnet* package for ANN than the LM and LMM models. However, there were no differences in accuracy between the ANNs and LM/LMM models developed for the thigh-worn accelerometer; moreover, the LM and LMM models for the thigh-worn accelerometer had similar accuracy to the ANNs for the hip-worn accelerometer and better accuracy for EE prediction than the LM and LMM models for the hip-worn accelerometer and both all models tested for the wrist-worn accelerometers.

The fact that the hip- and thigh-worn accelerometers experienced minimal, if any, loss in EE prediction accuracy with LM and LMM models compared to the ANNs is informative. LM and LMM models are much easier to implement and understand by researchers or clinicians who are not experts in the field of PA measurement, whereas machine learning models are considerably more complex to work with and understand. The complexity of machine learning models has resulted in reluctance to adopt them for use in PA assessment (Montoye *et al* 2016b), even though over a decade of research now supports machine learning as a viable approach for analyzing accelerometer data (Bao and Intille 2004, Parkka *et al* 2006, Preece *et al* 2009). Our finding that LM and LMM models, when developed to incorporate features extracted from raw hip- or thigh-accelerometer data, can be used with a high degree of accuracy for EE prediction provides a promising path forward for researchers using accelerometers who do not have access to machine learning techniques for analyzing accelerometer data.

The recent move to using raw data has allowed for higher accuracy to be achieved for wrist-worn accelerometers using both cut-point approaches (Esliger *et al* 2011) and machine learning models (Ellis *et al* 2014, Staudenmayer *et al* 2015, Strath *et al* 2015) compared to traditional LM modeling approaches. Given the use of wrist-worn accelerometers in several large studies including the NHANES 2011–2014 data collection cycle and the UK Biobank study, studies such as the current study are important as researchers work to develop ways of analyzing wrist-worn accelerometer data. We found that for the wrist-worn accelerometers, the ANN models had considerably better accuracy than the LM and LMM models. Being a more distal placement site than at the hip or thigh, the wrist likely experiences much greater and more variable accelerations than the hip or thigh, which lie closer to the body's center of mass. This added variability at the wrist site contributed to poor accuracy of PA measurement in early studies using wrist-worn activity monitors (Montoye *et al* 1983, Swartz *et al* 2000) and appears to have contributed to lower EE prediction accuracy in the current study. Another finding of this study was that there appeared little difference in predictive accuracy between the left and right wrists, although means for RMSE and correlations non-significantly trended toward slightly better accuracy of the left wrist-worn accelerometer; for most individuals in this study, this was the non-dominant wrist. This finding may help to inform choice of wrist placement when using accelerometers in field-based research.

A previous study by our research group found that adding more features extracted from the raw acceleration signal improved the accuracy of EE prediction from wrist-worn

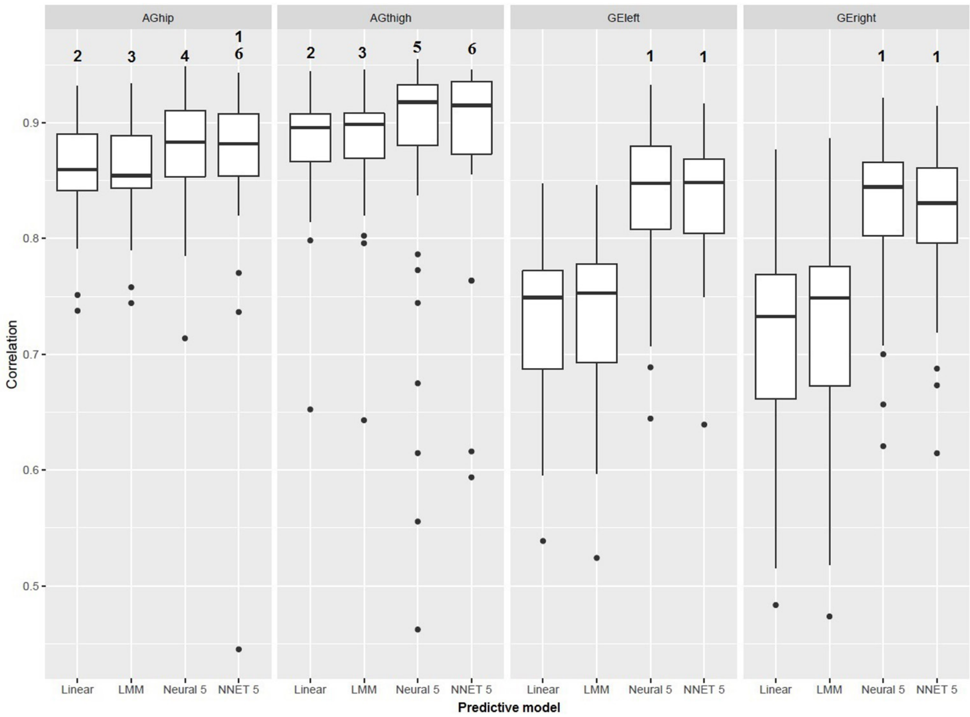

Figure 1. Correlations with measured energy expenditure for all accelerometer and model combinations. Data are displayed in box plots with median and inter-quartile range. Black dots represent outliers. (1) Indicates significant differences from linear regression and linear mixed models within the same accelerometer. (2) Indicates significant differences from all other accelerometers with the linear regression model. (3) Indicates significant differences from all other accelerometers with the linear mixed model. (4) Indicates significant differences from right wrist-worn accelerometer with the *nnet* ANN model. (5) Indicates significant differences from both wrist-worn accelerometer with the *nnet* ANN model. (6) Indicates significant differences from wrist-worn accelerometers with the *neuralnet* ANN model. ANN: artificial neural network. AGhip: ActiGraph accelerometer worn on hip. AGthigh: ActiGraph accelerometer worn on thigh. GEleft: GENEA accelerometer worn on left wrist. GEright: GENEA accelerometer worn on right wrist. Linear: linear regression model. LMM: linear mixed model. Neural 5: artificial neural network created with *neuralnet* package in R, with 5 units in hidden layer. NNET 5: artificial neural network created with *nnet* package in R, with 5 units in hidden layer.

accelerometers (Montoye *et al* 2015). Similarly, a previous analysis by Staudenmayer *et al* indicate that incorporation of more complex, frequency-domain features to machine learning models may also help achieve high EE prediction accuracy from wrist-worn accelerometers (Staudenmayer *et al* 2015). In contrast, our previous study found only a small improvement in accuracy of a hip-worn accelerometer and no improvement in accuracy of a thigh-worn accelerometer with added features to a machine learning model (Montoye *et al* 2015); a similar finding is reported by Kate *et al* who found that additional features did not improve EE prediction accuracy of a machine learning model developed for a hip-worn accelerometer (Kate *et al* 2016). Despite demonstrated advantages of wrist-worn accelerometers for improving wear compliance (Troiano *et al* 2014, Fairclough *et al* 2016) and assessment of variables such as activity type (Zhang *et al* 2012, Mannini *et al* 2013, Dong *et al* 2013, Kate *et al* 2016),

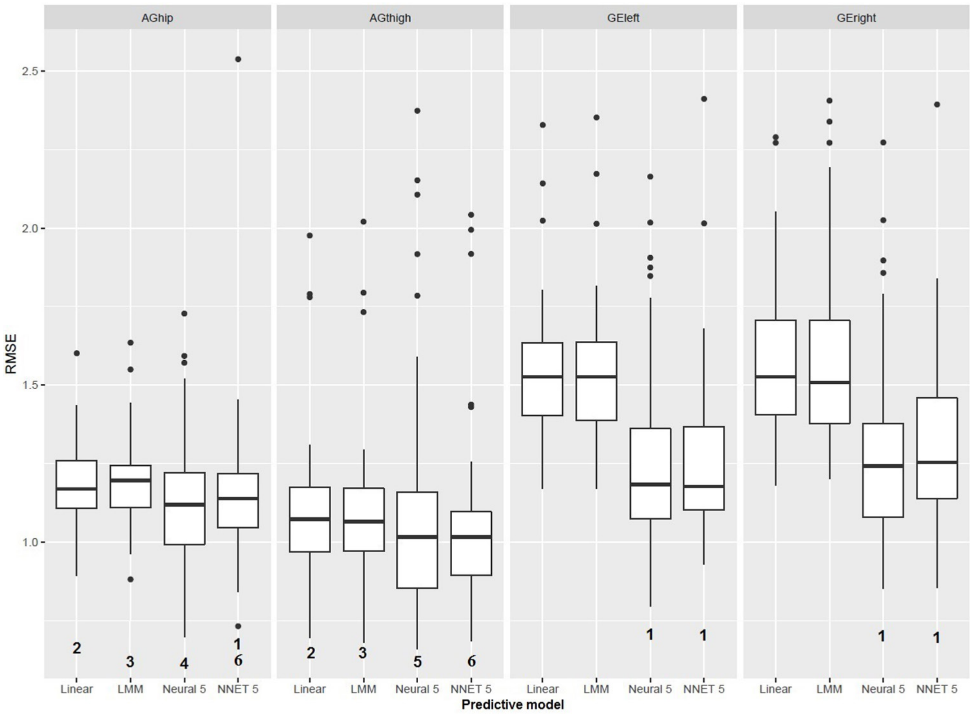

Figure 2. Root mean square error for predicted versus measured energy expenditure for all accelerometer and model combinations. Data are displayed in box plots with median and inter-quartile range. Black dots represent outliers. (1) Indicates significant differences from linear regression and linear mixed models within the same accelerometer. (2) Indicates significant differences from wrist-worn accelerometers with the linear regression model. (3) Indicates significant differences from wrist-worn accelerometers with the linear mixed model. (4) Indicates significant differences from right wrist-worn accelerometer with the *nnet* ANN model. (5) Indicates significant differences from both wrist-worn accelerometer with the *nnet* ANN model. (6) Indicates significant differences from wrist-worn accelerometers with the *neuralnet* ANN model. RMSE: root mean square in METs. ANN: artificial neural network. AGhip: ActiGraph accelerometer worn on hip. AGthigh: ActiGraph accelerometer worn on thigh. GEleft: GENEA accelerometer worn on left wrist. GEright: GENEA accelerometer worn on right wrist. Linear: Linear regression model. LMM: Linear mixed model. Neural 5: artificial neural network created with *neuralnet* package in R, with 5 units in hidden layer. NNET 5: artificial neural network created with *nnet* package in R, with 5 units in hidden layer.

sedentary behavior (Rowlands *et al* 2014, Rowlands *et al* 2015), and sleep (van Hees *et al* 2015, Buman *et al* 2016, Kamper *et al* 2016,), the trade-off for these advantages is that more complex modeling methods may be needed to optimize EE prediction accuracy of wrist-worn accelerometers, whereas simpler modeling can be used with high accuracy for hip- or thigh-worn accelerometers.

Another finding of our study is that there were minimal differences in EE prediction accuracy of ANN models fitted by two different packages (*nnet* and *neuralnet*) in R, which is expected. The inherent modeling scheme is essentially the same under both *nnet* and *neuralnet* packages. However, it is to be noted that the *nnet* is an older package in R and is simpler to use; the *nnet* package has also been more commonly used in previous work, with

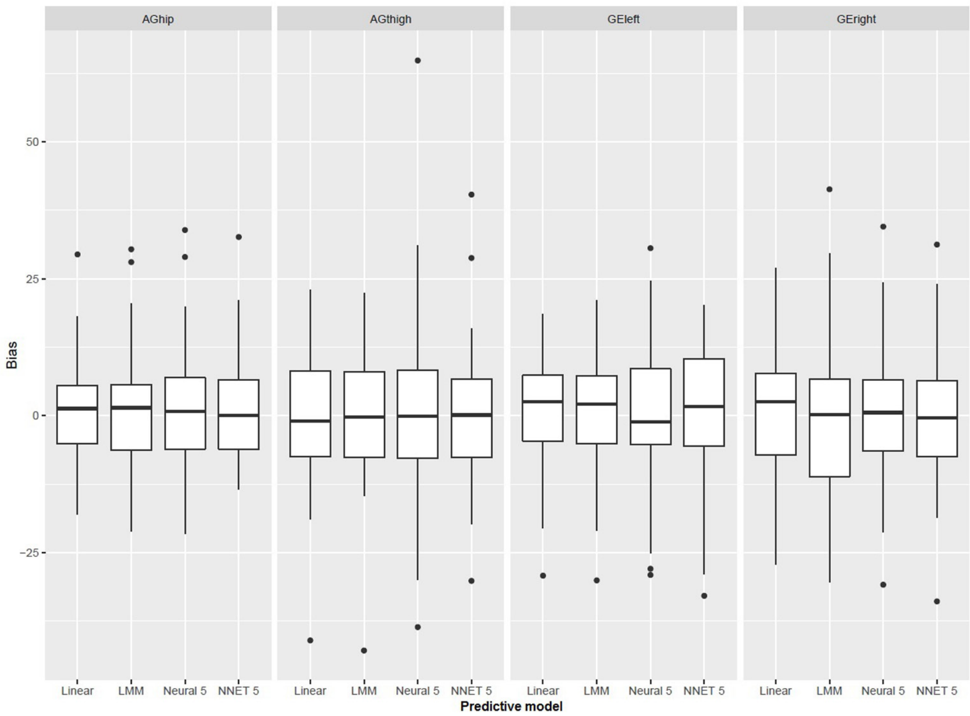

Figure 3. Bias for predicted versus measured energy expenditure for all accelerometer and model combinations. Data are displayed in box plots with median and inter-quartile range. Black dots represent outliers. ANN: artificial neural network. AGhip: ActiGraph accelerometer worn on hip. AGthigh: ActiGraph accelerometer worn on thigh. GEleft: GENEA accelerometer worn on left wrist. GEright: GENEA accelerometer worn on right wrist. Linear: Linear regression model. LMM: Linear mixed model. Neural 5: artificial neural network created with *neuralnet* package in R, with 5 units in hidden layer. NNET 5: artificial neural network created with *nnet* package in R, with 5 units in hidden layer.

the Staudenmayer *et al* study in 2009 providing code for development of ANN models in R (Staudenmayer *et al* 2009, Montoye *et al* 2015, Staudenmayer *et al* 2015). Therefore, we recommend using the *nnet* package for researchers interested in using or developing ANN models for PA assessment.

Finally, to assess performance of the models developed in this study, we compared their accuracy to that of the Freedson 1998 MET prediction equation, which is among the earliest and most widely used EE prediction equations developed and often used as a standard in the field (Freedson *et al* 1998). Data from the Freedson equation for this study were previously reported in our previous work (Montoye *et al* 2015). The LM and LMM models developed for the wrist-worn accelerometers had significantly lower correlations with measured EE than the Freedson equation, although RMSE was not different among the Freedson equation and LM/LMM models for the wrist-worn accelerometers. Conversely, RMSE was significantly lower, and correlations significantly higher, for all ANN models for all four accelerometers compared to the Freedson equation, (only exception was no difference between the correlation for the *neuralnet* ANN for the right wrist-worn accelerometer and the correlation for the Freedson equation). Previous work has shown that machine learning models developed to analyze triaxial, raw acceleration data from accelerometers have superior accuracy to traditional

LM approaches; the current study extends these findings, providing evidence that linear models developed to analyze raw data from hip- or thigh-worn accelerometers represent a viable approach to improving accelerometer EE prediction accuracy without needing to use complex, machine learning techniques.

Study strengths include the semi-structured nature of the activity protocol, numerous accelerometer locations, and high-quality criterion measure for EE. Additionally, raw data from GENEA and ActiGraph accelerometers have been shown to be comparable in past work and provide equally high accuracy for PA assessment, providing confidence that our results were not influenced by the use of two different accelerometer brands (John *et al* 2013, Hildebrand *et al* 2014). Study limitations include the homogenous sample of young, relatively lean participants, which may limit applicability of the models created to a more diverse or older population. Additionally, the activity protocol utilized in this study resulted in participants spending a low proportion of the study in sedentary behaviors and does not reflect the lifestyle patterns of most adults, who spend the majority of waking hours in sedentary behaviors (Matthews *et al* 2008, Donaldson *et al* 2015). Further validation in a free-living setting may be necessary. Finally, it is worth noting that there were many outlying data points, especially for correlations and RMSE (figures 1 and 2), which most often were in the direction of poorer measurement accuracy for some of the study participants. It is common that predictive models poorly assess some individuals, and this should be considered when employing predictive models for field-based research.

Conclusions

Our study indicates that ANN models developed to predict EE from features extracted from raw acceleration data can significantly improve prediction accuracy for wrist-worn accelerometers but may yield minimal improvements in accuracy from hip- or thigh-worn accelerometers, compared to LM and LMM models. Both LM and LMM models developed to predict EE from hip- or thigh-worn accelerometers had high EE prediction accuracy and may be simpler alternatives to using machine learning for EE prediction from accelerometer data.

Acknowledgments

The authors thank Todd Buckingham, Ryan Hulteen, Brandon Krinock, Elizabeth Lenz, and Stefan Rowland for their assistance in data collection and participant recruitment. This research was supported by the Michigan State University College of Education and by a Student Award Program grant from the Blue Cross Blue Shield of Michigan Foundation (1997.SAP).

References

Akkermans M A, Sillen M J, Wouters E F and Spruit M A 2012 Validation of the oxycon Mobile metabolic system in healthy subjects *J. Sports Sci. Med.* **11** 182–3

Bao L and Intille S S 2004 Activity recognition from user-annotated acceleration data *Proc. Pervasive* **3001** 1–17

Buman M P, Hu F, Newman E, Smeaton A F and Epstein D R 2016 Behavioral periodicity detection from 24 h wrist accelerometry and associations with cardiometabolic risk and health-related quality of life *Biomed. Res. Int.* 4856506

Cleland I, Kikhia B, Nugent C, Boytsov A, Hallberg J, Synnes K, McClean S and Finlay D 2013 Optimal placement of accelerometers for the detection of everyday activities *Sensors* **13** 9183–200

Crouter S E, Clowers K G and Bassett D R 2006 A novel method for using accelerometer data to predict energy expenditure *J. Appl. Physiol.* **100** 1324–31
Donaldson S C, Montoye A H, Tuttle M S and Kaminsky K A 2015 Variability of objectively measured sedentary behavior *Med. Sci. Sports Exerc.* **48** 755–61
Dong B, Montoye A, Moore R, Pfeiffer K and Biswas S 2013 Energy-aware activity classification using wearable sensor networks *Proc. SPIE* 87230y
Ellis K, Kerr J, Godbole S, Lanckriet G, Wing D and Marshall S 2014 A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers *Physiol. Meas.* **35** 2191–203
Esliger D W, Rowlands A V, Hurst T L, Catt M, Nurray P and Eston R G 2011 Validation of the GENEA accelerometer *Med. Sci. Sports Exerc.* **43** 1085–93
Fairclough S J, Noonan R, Rowlands A V, Van Hees V, Knowles Z and Boddy L M 2016 Wear compliance and activity in children wearing wrist- and hip-mounted accelerometers *Med. Sci. Sports Exerc.* **48** 245–53
Freedson P S, Melanson E and Sirard J 1998 Calibration of the Computer Science and Applications, Inc. accelerometer *Med. Sci. Sports Exerc.* **30** 777–81
Glickman M E, Rao S R and Schultz M R 2014 False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies *J. Clin. Epidemiol.* **67** 850–7
Gunther F and Fritch S 2010 Neuralnet: training of neural networks *R J.* **2** 30–8
Ham S A, Reis J P, Strath S J, Dubose K D and Ainsworth B E 2007 Discrepancies between methods of identifying objectively determined physical activity *Med. Sci. Sports Exerc.* **39** 52–8
Hendelman D, Miller K, Baggett C, Debold E and Freedson P 2000 Validity of accelerometry for the assessment of moderate intensity physical activity in the field *Med. Sci. Sports Exerc.* **32** s442–9
Hildebrand M, Van Hees V T, Hansen B H and Ekelund U 2014 Age group comparability of raw accelerometer output from wrist- and hip-worn monitors *Med. Sci. Sports Exerc.* **46** 1816–24
John D D and Freedson P 2012 ActiGraph and actical physical activity monitors: a peek under the hood *Med. Sci. Sports Exerc.* **44** s86–9
John D, Sasaki J, Staudenmayer J, Mavilia M and Freedson P S 2013 Comparison of raw acceleration from the GENEA and ActiGraph GT3X + activity monitors *Sensors* **13** 14754–63
Kamper J E, Garofano J, Schwartz D J, Silva M A, Zeitzer J, Modarres M, Barnett S D and Nakase-Richardson R 2016 Concordance of actigraphy with polysomnography in traumatic brain injury neurorehabilitation admissions *J. Head Trauma Rehabil.* **31** 117–25
Kate R J, Swartz A M, Welch W A and Strath S J 2016 Comparative evaluation of features and techniques for identifying activity type and estimating energy cost from accelerometer data *Physiol. Meas.* **37** 360–79
Lyden K, Keadle S K, Staudenmayer J and Freedson P S 2014 A method to estimate free-living active and sedentary behavior from an accelerometer *Med. Sci. Sports Exerc.* **46** 386–97
Malina R 1995 *Anthropometry* (*Physiological Assessment of Human Fitness*) (Champaign, IL: Human Kinetics, Inc) pp 205–19
Mannini A, Intille S S, Rosenberger M, Sabatini A M and Haskell W 2013 Activity recognition using a single accelerometer placed at the wrist or ankle *Med. Sci. Sports Exerc.* **45** 2193–203
Mannini A and Sabatini A M 2010 Machine learning methods for classifying human physical activity from on-body accelerometers *Sensors* **10** 1154–75
Matthews C E, Chen K Y, Freedson P S, Buchowski M S, Beech B M, Pate R R and Troiano R P 2008 Amount of time spent in sedentary behaviors in the United States, 2003–2004 *Am. J. Epidemiol.* **167** 875–81
Montoye A H, Dong B, Biswas S and Pfeiffer K A 2016a Validation of a wireless accelerometer network for energy expenditure measurement *J. Sports Sci.* **34** 2130–9
Montoye A H, Moore R W, Bowles H R, Korycinski R and Pfeiffer K A 2016b Reporting accelerometer methods in physical activity intervention studies: a systematic review and recommendations for authors *Br. J. Sports Med.* (doi: 10.1136/bjsports-2015-095947)
Montoye A H, Mudd L M, Biswas S and Pfeiffer K A 2015 Energy expenditure prediction using raw accelerometer data in simulated free living *Med. Sci. Sports Exerc.* **47** 1735–46
Montoye A H K, Pivarnik J M, Mudd L M, Biswas S and Pfeiffer K A 2016c Wrist-independent energy expenditure prediction models from raw accelerometer data *Physiol. Meas.* **37** 1770–84
Montoye H J, Washburn R, Servais S, Ertl A, Webster J G and Nagle F J 1983 Estimation of energy expenditure by a portable accelerometer *Med. Sci. Sports Exerc.* **15** 403–7

Parkka J, Ermes M, Korpipaa P, Mantyjarvi J, Peltola J and Korhonen I 2006 Activity classification using realistic data from wearable sensors *IEEE Trans. Inf. Technol. Biomed.* **10** 119–28
Preece S J, Goulermas J Y, Kenney L P, Howard D, Meijer K and Crompton R 2009 Activity identification using body-mounted sensors—a review of classification techniques *Physiol. Meas.* **30** r1–33
R Core Development Team R: a language and environment for statistical computing Version 2.12.1 (available: www.r-project.org/)
Rosdahl H, Gullstrand L, Salier-Eriksson J, Johansson P and Schantz P 2010 Evaluation of the Oxycon Mobile metabolic system against the Douglas bag method *Eur. J. Appl. Physiol.* **109** 159–71
Rowlands A V, Olds T S, Hillsdon M, Pulsford R, Hurst T L, Eston R G, Gomersall S R, Johnston K and Langford J 2014 Assessing sedentary behavior with the GENEActiv: introducing the sedentary sphere *Med. Sci. Sports Exerc.* **46** 1235–47
Rowlands A V, Yates T, Olds T S, Davies M, Khunti K and Edwardson C L 2015 Sedentary sphere: wrist-worn accelerometer-brand independent posture classification *Med. Sci. Sports Exerc.* **48** 748–54
Staudenmayer J, He S, Hickey A, Sasaki J and Freedson P 2015 Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements *J. Appl. Physiol.* **119** 396–403
Staudenmayer J, Pober D, Crouter S, Bassett D and Freedson P 2009 An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer *J. Appl. Physiol.* **107** 1300–7
Strath S J, Bassett D R and Swartz A M 2003 Comparison of MTI accelerometer cut-points for predicting time spent in physical activity *Int. J. Sports Med.* **24** 298–303
Strath S J, Kate R J, Keenan K G, Welch W A and Swartz A M 2015 Ngram time series model to predict activity type and energy cost from wrist, hip and ankle accelerometers: implications of age *Physiol. Meas.* **36** 2335–51
Swartz A M, Strath S J, Bassett D R, O'Brien W L, King G A and Ainsworth B E 2000 Estimation of energy expenditure using csa accelerometers at hip and wrist sites *Med. Sci. Sports Exerc.* **32** s450–6
Troiano R P, McClain J J, Brychta R J and Chen K Y 2014 Evolution of accelerometer methods for physical activity research *Br. J. Sports Med.* **48** 1019–23
Trost S G, Wong W K, Pfeiffer K A and Zheng Y 2012 Artificial neural networks to predict activity type and energy expenditure in youth *Med. Sci. Sports Exerc.* **44** 1801–9
van Hees V T, Sabia S, Anderson K N, Denton S J, Oliver J, Catt M, Abell J G, Kivimaki M, Trenell M I and Singh-Manoux A 2015 A novel, open access method to assess sleep duration using a wrist-worn accelerometer *PLoS One* **10** e0142533
Vathsangam H, Emken A, Schroeder E T, Spruijt-Metz S and Sukhatme G S 2011 Determining energy expenditure from treadmill walking using hip-worn inertial sensors: an experimental study *IEEE Trans. Biomed. Eng.* **58** 2804–15
Vathsangam H, Emken A, Schroeder E T, Spruijt-Metz S and Sukhatme G S 2010 Energy estimation of treadmill walking using on-body accelerometers and gyroscopes *Conf. Proc. Engineering in Medicine and Biology Society* **2010** 6497–501
Zhang S, Rowlands A V, Murray P and Hurst T L 2012 Physical activity classification using the GENEA wrist-worn accelerometer *Med. Sci. Sports Exerc.* **44** 742–8