PAPER

Lab-based validation of different data processing methods for wrist-worn ActiGraph accelerometers in young adults

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

Physiol. Meas. **38** (2017) 1045–1060 https://doi.org/10.1088/1361-6579/aa6d00

Lab-based validation of different data processing methods for wrist-worn ActiGraph accelerometers in young adults

Laura D Ellingson[1,6], Paul R Hibbing[1,2], Youngwon Kim[3], Laura A Frey-Law[4], Pedro F Saint-Maurice[1,5] and Gregory J Welk[1]

[1] Department of Kinesiology, Iowa State University, Ames Iowa, United States of America
[2] Department of Kinesiology, Recreation, and Sport Studies, University of Tennessee, Knoxville TN, United States of America
[3] MRC Epidemiology Unit, University of Cambridge, Cambridge, United Kingdom
[4] Department of Physical Therapy and Rehabilitation Science, University of Iowa, Iowa City, IA, United States of America
[5] School of Psychology, University of Minho, Braga, Portugal

E-mail: ellingl@iastate.edu

Received 19 January 2017, revised 6 April 2017
Accepted for publication 12 April 2017
Published 8 May 2017

Abstract

The wrist is increasingly being used as the preferred site for objectively assessing physical activity but the relative accuracy of processing methods for wrist data has not been determined. *Objective:* This study evaluates the validity of four processing methods for wrist-worn ActiGraph (AG) data against energy expenditure (EE) measured using a portable metabolic analyzer (OM; Oxycon mobile) and the Compendium of physical activity. *Approach:* Fifty-one adults (ages 18–40) completed 15 activities ranging from sedentary to vigorous in a laboratory setting while wearing an AG and the OM. Estimates of EE and categorization of activity intensity were obtained from the AG using a linear method based on Hildebrand cutpoints (HLM), a non-linear modification of this method (HNLM), and two methods developed by Staudenmayer based on a Linear Model (SLM) and using random forest (SRF). Estimated EE and classification accuracy were compared to the OM and Compendium using Bland–Altman plots, equivalence testing, mean absolute percent error (MAPE), and Kappa statistics. *Main results:* Overall, classification agreement with the Compendium was similar across methods ranging from a Kappa of 0.46 (HLM) to 0.54 (HNLM). However, specificity

[6] Department of Kinesiology, Iowa State University, 239 Forker, Ames, IA 50011, United States of America.

and sensitivity varied by method and intensity, ranging from a sensitivity of 0% (HLM for sedentary) to a specificity of ~99% for all methods for vigorous. None of the methods was significantly equivalent to the OM ($p > 0.05$). *Significance:* Across activities, none of the methods evaluated had a high level of agreement with criterion measures. Additional research is needed to further refine the accuracy of processing wrist-worn accelerometer data.

Keywords: accelerometer, wrist, measurement, objective, physical activity, sedentary

S Supplementary material for this article is available online

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

Introduction

A variety of accelerometer-based processing methods are currently available for analyzing objectively measured physical activity (PA) data (Troiano *et al* 2014), but there is little consensus about the most appropriate monitors and methods. The challenges in standardization have been further compounded by the increased emphasis on wrist-worn monitors in consumer and research applications—including adoption in several large epidemiological studies (e.g. NHANES, UK Biobank) (Freedson and John 2013, UKBiobank 2016). Insights and applications developed for hip-worn monitors do not readily translate to wrist-worn monitors. With this, a systematic evaluation of different monitors and processing methods is needed to advance and standardize research.

The transition to wrist-worn monitors has coincided with an increased emphasis on developing methods using raw acceleration data, as opposed to using older methods based on the more traditional 'activity counts'. One goal of this transition is obtaining comparable data from different brands and types of monitors. Raw acceleration data from wrist-worn monitors have been used to characterize activity patterns in large cohorts of individuals (Da Silva *et al* 2014, Van Hees *et al* 2014), but estimates may not be comparable to past results. Standardized methods and activity cutpoints have been proposed based on raw wrist acceleration data (Hildebrand *et al* 2014) but they have not been cross-validated in independent samples. There are also several new statistical and pattern-recognition approaches proposed for processing raw acceleration data at the wrist (Lyden *et al* 2014, Staudenmayer *et al* 2015). These methods may offer unique advantages for assessing PA. However, direct comparisons are needed to identify the best methods as well as to determine the implications of this recent transition in accelerometer data processing/methodology.

Therefore, the primary goal of this study was to test the relative accuracy of these new methods for processing wrist-worn accelerometer data against laboratory-based measures of energy expenditure (EE) across a range of intensities from sedentary to vigorous. Consistent with best practices (Bassett *et al* 2012) and our past validation studies (Kim and Welk 2015, Ellingson *et al* 2016), we compared the accuracy of four processing methods with measured METs and the Compendium of PA (Ainsworth *et al* 2011). This is a secondary data analysis using the same raw dataset utilized in the two recently published validation studies from our group noted above (Kim and Welk 2015, Ellingson *et al* 2016).

Materials and methods

Participants

Participants were 51 adults (57% male; average ± SD age 23.5 ± 4.6 years) with an average body mass index of 23.3 ± 3.4 kg m^{-2}. Participants were recruited from the Iowa State University community via flyers and word of mouth. Exclusion criteria included metal allergies or having mobility limitations that precluded participation in PA.

Procedures

Study procedures were approved by the Institutional Review Board and participants read and signed informed consent documents prior to data collection. Data collection procedures are thoroughly detailed in our previous publication (Kim and Welk 2015). Briefly, all participants performed a series of 15 common activities in a controlled laboratory setting. Activities were selected to represent the four intensity categories (sedentary, light, moderate, vigorous) and a mixture of tasks that involve a range of wrist movements: supine resting, sitting reading a book, sitting typing, sitting fidgeting, standing reading a book, standing typing, standing fidgeting, climbing stairs, throwing/catching a ball, stationary biking, walking on a treadmill at 2 mph and 3 mph, walking at 3 mph typing, and running on a treadmill at 4.5 and 5.5 mph. Activities were performed in the order listed above for 5 min each, with 1 min resting intervals separating activities. While completing the activities, the participants concurrently wore a wrist-mounted ActiGraph GT3X+ accelerometer and a portable indirect calorimetry system (Oxycon-Mobile; OM), both detailed below.

The ActiGraph GT3X+ (AG; ActiGraph LLC, Pensacola, FL, USA) is a small (4.6 × 3.3 × 1.5 cm), light-weight (19 g) tri-axial accelerometer that records acceleration ranging from −6 to 6 g. It samples acceleration at a selected rate between 30–100 Hz, which is then digitized through a 12-bit analog-to-digital converter. The digital values are filtered via a band-pass filter at a range between 0.24–2.5 Hz. In the present study, the GT3X+ was placed on the right wrist and initialized at 100 Hz. Appropriate placement for the monitor was ensured by a member of the research team. Data were recorded continuously during the 15 activities performed in the lab.

Criterion measures

Indirect calorimetry was used as the criterion for estimates of EE. The Oxycon mobile (OM; CareFusion Corp, San Diego, CA) is a portable indirect calorimetry system that measures breath-by-breath respiratory gas exchange under laboratory settings as well as free-living environments. Previous research has demonstrated the validity and reliability of the OM (Hodges *et al* 2005, Salier Eriksson *et al* 2012), and it is commonly used as a criterion method to provide accurate values of EE. Prior to each test, gas and volume calibrations were performed according to manufacturer's recommendations. The flowmeter was calibrated automatically. Acceptable percent error ranges for gas calibration and volume calibration were 3% and 2%, respectively. Oxygen consumption data from the OM were downloaded in ml/kg/min for each minute of testing and averaged for each of the 15 activities.

In addition to the OM criterion for EE, the Compendium of PA (Ainsworth *et al* 2011) was used as the criterion for intensity classification (sedentary, light, moderate, and vigorous) in order to standardize the comparisons between participants. This is an appropriate criterion for observed activities (particularly for sedentary and light activities that are differentiated by

posture and lack of movement). Using the Compendium as a criterion simplified the categorical analysis and made the outcomes more generalizable. Participants were observed by two research assistants during all laboratory procedures and notes were taken throughout each session to document that each activity was performed as prescribed.

Data processing

Data were processed using four different methods of varying complexity to enable direct comparisons of outcomes relative to the criterion measure as well as the variability between outcomes. The two simpler methods utilize the acceleration processing methods described by Hildebrand *et al* (2014), relying on the resultant magnitude of the triaxial accelertions minus gravity (Euclidean Norm Minus One; ENMO) and the two more complex methods were drawn from the study by Staudenmayer *et al* relying on variation and angle of resultant accelerations or a decision tree (Staudenmayer *et al* 2015).

Hildebrand models

The Hildebrand linear method (HLM) estimates METs from raw acceleration data, collected at 100 Hz. The resultant movement accelerations (g's) are calculated as the vector magnitude −1 to account for gravity over each second. Second-by-second estimates of VO_2 (ml/kg/min) were then generated using equation (1) developed by Hildebrand and colleagues (Hildebrand *et al* 2014) for adults wearing the monitor at the non-dominant wrist, then dividing the result by 3.5 ml/kg/min to obtain the EE in $MET_{3.5}$, as shown in equation (2). Data files were downloaded in comma separated values (CSV) format and processed using a function provided by Hildebrand to calculate the ENMO acceleration for epochs of user-defined length using R (Anon 2016).

$$VO_2 = 0.0320\,(\text{ENMO}) + 7.28, \tag{1}$$

where ENMO is acceleration in millig's

$$MET_{3.5} = VO_2/3.5. \tag{2}$$

The Hildebrand nonlinear method (HNLM) uses the same processing method described above to calculate the resultant wrist accelerations, but then uses a modified, non-linear equation to estimate EE. It was developed by fitting Hildebrand's reported data, collectively with an entirely separate dataset, to a nonlinear model when it was recognized that the standard Hildebrand equation is unable to produce EE estimates below the intercept value of the linear equation (i.e. 7.28 ml/kg/min or 2.08 METs). The resulting power equation estimates VO_2 in ml/kg/min (equation (3)), which can be converted to METs using equation (2). As the non-linear model does not have an intercept inherent to the linear model (see equation (2)), a minimum VO_2 value of 3.0 was applied (floor value). Thus this nonlinear approach relies on the same acceleration parameter, ENMO (milig's), but uses a more complex model. See supplementary materials (stacks.iop.org/PM/38/1045/mmedia) for a more complete description of this method and the data collection and analysis procedures that were used to create the modified equation.

$$VO_2 = 0.901(\text{ENMO})^{0.534} \tag{3}$$

Staudenmayer models

The Staudenmayer methods (Staudenmayer *et al* 2015) use raw acceleration data to estimate PA intensity (including EE) using a linear model and random forest methods.

The Staudenmayer linear method (SLM) increases complexity over the Hildebrand approaches as it consists of a linear multiple regression equation based on two acceleration parameters. Several possible parameters were pairwise iterated, with the best performance being shown in the combination of the standard deviation of the vector magnitude (sdvm) from samples in 15 s windows and the mean acceleration angle (mangle) over the same period (equation (4)).

$$\text{METs} = 1.893\,78 + 5.508\,21\,(\text{sdvm}) - 0.027\,05\,(\text{mangle}) \tag{4}$$

The most complex approach evaluated here is Staudenmayer's random forest (SRF) method, which relies on probabilities assigned to leaf nodes of 500 decision trees created from random subsamples of the summary variables for each window. These summaries include the mean and standard deviation of the vector magnitude (not subtracting gravity) and angle of acceleration as well as features derived from fast fourier transform analysis of the signal. The model was developed using the R package random forest (Liaw and Wiener 2002) and more information is available in the package documentation, as well as in previous publications (Ellis *et al* 2014, Staudenmayer *et al* 2015).

The SLM and SRF were processed concurrently. For each raw acceleration file generated using the ActiLife software the following steps were performed: (1) read in the file, (2) compute the statistical features for each window and (3) enter the appropriate features into each model and return the results. After steps (1)–(3) were completed for each file, (4) the data were merged into a complete data set. The validation data used in the original paper were sampled at 80 Hz and examined in 15 s windows, with statistical features of the data distribution in these windows used as inputs to the models. We made minor and non-consequential modifications to the code to accommodate the fact that our data were collected at 100 Hz, as opposed to 80 Hz in the original paper (code available on request).

Following these initial processing steps, for each method the first and last minutes of each activity were removed (as well as the 1 min resting intervals between activities) to exclude extraneous data captured when transitioning between activities. MET estimates for minutes 2–4 of each activity were then averaged and used for comparisons between processing methods and criterion methods (OM, Compendium estimates). For comparisons of intensity classification, estimates of EE for each activity from each processing method were categorized into sedentary, light, moderate, or vigorous using standard cut-offs (sedentary: <1.5; light: 1.5–2.9; moderate: 3.0–5.9; vigorous: $\geqslant 6.0$) and compared with categorization of activities using the estimates of EE found in the Compendium of PA (Ainsworth *et al* 2011). The Compendium was used as the criterion here, because of individual variability in the OM-measured METs for sedentary and light activities. For example, there were cases where the MET value measured by the OM was either just over the cut-off for sedentary (e.g. 1.56 METs) or just under the cut-off for light intensity (e.g. 1.45 METs). There is consensus in the literature that sedentary behavior should be categorized both by METS and by posture so the use of the Compendium provided a way to capture the EE costs of the intended behavioral category (rather than relying exclusively on the EE values). See supplementary materials for a list of the activities and their measured intensity from the OM).

Statistical analyses

This study evaluated agreement of the four methods described above against criterion measures from indirect calorimetry and The Compendium. As in our previous validation papers (Kim and Welk 2015, Ellingson *et al* 2016), for each method we adjusted the MET estimates based on individualized estimates of basal metabolic rate (BMR) using the prediction equation developed by Schofield (1985) in order to facilitate direct comparisons of EE with the OM and among the different processing methods. Thus, the MET_{BMR} for the OM were obtained by dividing measured VO_2 values (in ml/kg/min) by the estimated BMR value. For each of the four processing methods, MET_{BMR} was calculated by multiplying average MET values for each 3 min period by 3.5, then dividing by the estimated BMR values.

Descriptive analyses were performed to summarize MET_{BMR} values across the trial and by intensity for each processing method. Mean absolute percent error (MAPE) and Bland–Altman plots were used to characterize the accuracy of measures and make comparisons among the processing methods. Absolute errors (MAPE) provide the advantage of evaluating total error without over- and under-estimation errors cancelling out. Equivalence testing was also performed to determine the overall group-level agreement between the four processing techniques and the OM. This statistical test provides evidence of 'equivalence' by reversing the traditional null hypothesis of no difference to specify that methods compared are not equivalent to the criterion. As such, rejecting the null hypothesis indicates significant equivalence between the methods and the OM. We compared 90% confidence intervals (CI) of the means for each of the four different test methods at each activity intensity as well as for minutes of moderate and vigorous physical activity (MVPA) against a pre-specified zone of equivalence, which was defined as $\pm 10\%$ of the mean of the OM. If the 90% CIs for the methods completely fell within the equivalence zone, they were considered to be significantly equivalent to the OM. Finally, diagnostic tests (i.e. Kappa statistics, Sensitivity (Se), and Specificity (Sp)) were performed to evaluate classification agreement between the methods and Compendium across the different intensity categories. We also computed the absolute agreement per activity in order to understand the contribution of activity type to overall and intensity-specific misclassification rates.

Results

Descriptive analyses for EE estimates and summaries for MAPE are shown in table 1. When averaging across the whole trial, estimates of EE were similar ranging from a MET_{BMR} of 2.74 for HNLM to a MET_{BMR} of 3.48 for HLM. Overall MAPEs were relatively small across the trial, ranging from 9.32 (HLM) to 24.59 (HNLM). However, when looking at each intensity individually, results showed greater inconsistencies as compared with the OM. For sedentary and light intensity activities, estimates from the SRF, SLM, and HLM methods resulted in relatively large errors at the individual level, demonstrated by MAPEs ranging from ~35 to 76%. At these lower intensities, the HNLM provided estimates of EE that were much closer to the OM with a MAPE of ~23% for sedentary and ~19% for light intensity. For moderate intensity activities, all four methods produced relatively large errors, but this pattern was somewhat reversed in that SLM, and HLM method provided estimates of EE that were closest to the OM at ~36% MAPE while the SRF and HNLM methods had larger MAPEs of ~45–50%. Each of the processing methods performed relatively well for vigorous intensity with closer estimates of EE ranging from ~13% (HNLM & SRF) to ~15–20% (SRF & HLM).

The Bland–Altman plots shown in figure 1 further illustrate this variability across intensities and methods. For sedentary and light intensity activities, SRF, SLM, and HLM methods

Table 1. Indicators of agreement for energy expenditure (METBMR estimate) for the whole trial and by activity intensity for Hildebrand Linear (HLM) and Nonlinear Methods (HNLM) and Staudenmayer's random forest (SRF) and Linear Model (SLM), in relation to the Oxycon mobile (OM); MAPE: mean absolute percent error.

		MET_{BMR} estimate mean (SD)	MAPE mean (SD)
Whole trial	OM	3.63 (0.37)	NA
	HLM	3.48 (0.37)	9.32 (7.89)
	HNLM	2.74 (0.38)	24.59 (9.52)
	SLM	3.31 (0.72)	19.24 (12.97)
	SRF	3.18 (0.37)	14.60 (9.49)
Sedentary	OM	1.36 (0.20)	NA
	HLM	2.34 (0.19)	76.01 (31.49)
	HNLM	1.13 (0.26)	23.39 (11.91)
	SLM	2.01 (0.61)	57.26 (42.22)
	SRF	1.78 (0.36)	35.86 (30.56)
Light	OM	1.84 (0.34)	NA
	HLM	2.75 (0.32)	53.73 (28.91)
	HNLM	1.89 (0.45)	19.34 (17.91)
	SLM	2.85 (0.94)	62.88 (48.94)
	SRF	2.63 (0.50)	47.35 (31.81)
Moderate	OM	4.54 (0.61)	NA
	HLM	2.83 (0.23)	36.54 (10.57)
	HNLM	2.19 (0.35)	50.91 (10.74)
	SLM	2.77 (0.63)	37.95 (16.88)
	SRF	2.46 (0.41)	44.50 (13.16)
Vigorous	OM	9.56 (0.85)	NA
	HLM	8.81 (1.78)	16.75 (10.28)
	HNLM	8.78 (1.29)	13.45 (8.48)
	SLM	7.93 (1.46)	20.07 (12.74)
	SRF	8.36 (0.78)	13.73 (8.51)

tended to over-estimate EE by ~0.5–1 MET. In contrast, the HNLM had narrower limits of agreement than SRF or SLM, which centered around zero. Further, the SLM and SRF methods showed a systematic bias at these lower intensities characterized by a tendency to overestimate METs when OM estimates were higher ($r_{range} = 0.38$ to 0.81). For moderate intensity activities all methods tended to underestimate METs by approximately 1.5–2, with similar ranges for the limits of agreement. Both the HLM and HNLM methods showed more apparent patterns of systematic bias at this intensity characterized by a tendency to underestimate METs when measured METs were higher ($r_{range} = -0.50$ to -0.74). This was particularly true for the moderate intensity activities that did not involve a high degree of wrist motion relative to the intensity level (stationary cycling, walking while typing), whereas the SRF underestimated walking intensity, but provided more systematically accurate estimates of these other activities. Vigorous intensity activities were also generally underestimated by ~1.5–2 METs by each of the four processing methods. However, the SRF method had narrower limits of agreement than the other processing methods; HLM and HNLM methods showing the widest limits of agreement.

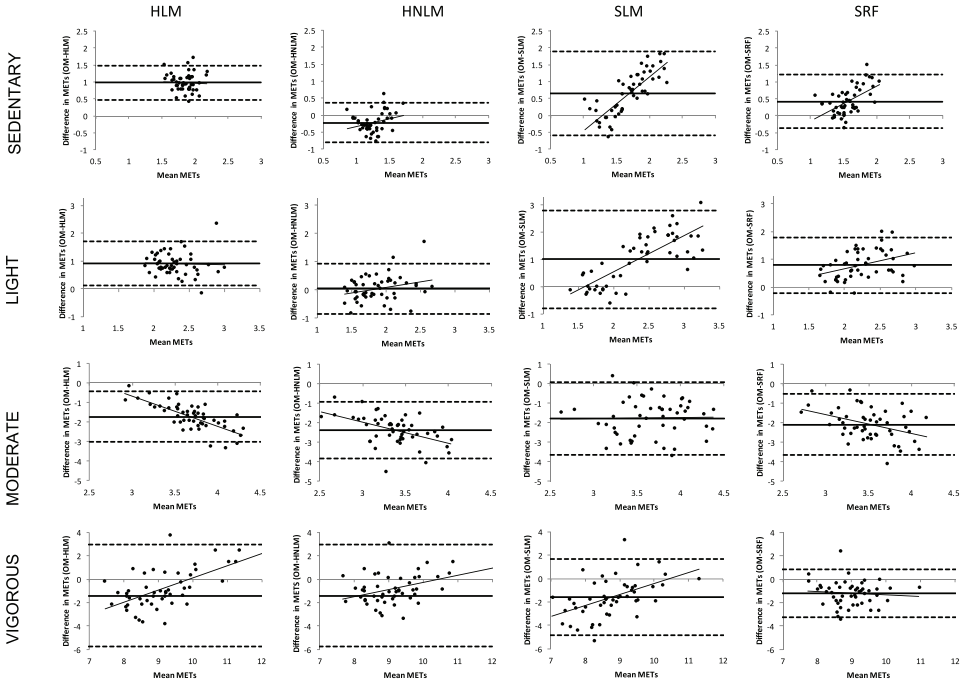

Figure 1. Bland–Altman Plots demonstrating comparisons of each processing method to the Oxycon mobile (OM) as the criterion for 4 levels of task intensity; HLM: Hildebrand linear method, HNLM: Hildebrand nonlinear method, SLM: Staudenmayer linear model, SRF: Staudenmayer random forest.

Results from the equivalence testing are illustrated in figure 2. Averaging across activity intensities, only the HLM method was significantly equivalent to the OM ($p < 0.05$). However, when looking at each intensity separately and MVPA, none of the four methods achieved significant equivalence with the OM ($p > 0.05$). For sedentary activities, the HNLM estimate for EE was centered around the MET measured by the OM, but the confidence intervals were wider than the OM's band of equivalency. Similarly for light intensity activity, the HNLM overlapped with the OM but the upper confidence interval was above that of the OM.

Classification performance for each of the methods at each intensity is shown in table 2. The overall accuracy was generally comparable among methods, with the HNLM and SRF methods showing the highest levels across intensities at 54.0% and 53.5%, respectively and with the highest Kappa statistics at 0.38 (HNLM) and 0.36 (SRF). However, when looking at each intensity individually, Se and Sp varied widely among methods. The HNLM method was notably superior to the other methods regarding Se for sedentary time (Se = 89.4%) and the Sp for light intensity activities (Sp = 90.3%). Sensitivity was generally poor across all processing methods for moderate intensity activities, ranging from 37 to 48%. However, three of the four moderate intensity activities involved disproportionately low levels of wrist motion (stair climbing, cycling, walking while typing). For vigorous intensity activities, both Se and Sp were quite good for all methods and ranged from 89 to 99%.

Our last set of analyses examined variability in classification agreement for each of the 15 activities. The percent of measurements that were accurately classified by intensity category for each activity and processing method is illustrated in figure 3. Across sedentary activities the HNLM method was superior to the other three methods, accurately classifying these activities

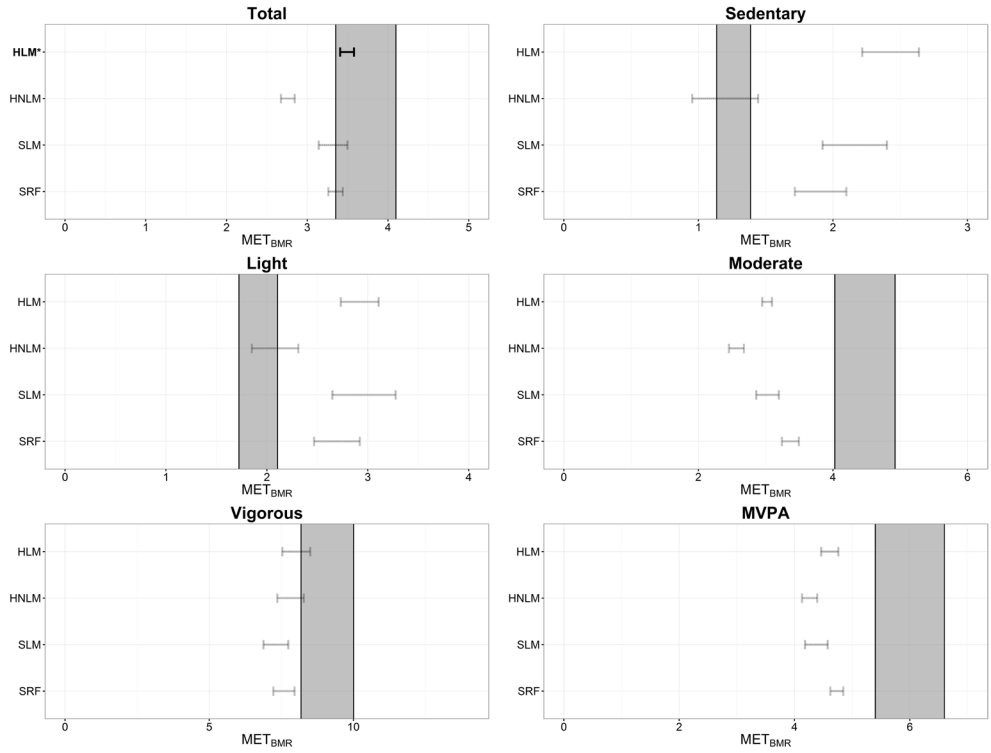

Figure 2. Analyses demonstrating overall equivalency between each processing method to Oxycon mobile (OM) measurement of energy intensity for 15 tasks classified as: total physical activity, sedentary, light, moderate, vigorous, and combined moderate and vigorous. The shaded gray area is the zone of equivalence (±10%), and the error bars are 90% CIs. HLM: Hildebrand Linear method, HNLM: Hildebrand Nonlinear method, SLM: Staudenmayer linear model, SRF: Staudenmayer random forest.

89.7% percent of the time. Notably, the SRF, SLM, and HLM tended to overestimate activity intensity particularly for sitting typing and for sitting fidgeting, perhaps due to greater involvement of hand movements in these activities as well as the floor value of the HLM methodology. For light intensity activities, the HLM method was superior to other methods with an overall accuracy of 67.3%. Within this category, all methods tended to overestimate the EE for throwing and catching a ball, likely due to the large arm movements that occur during this activity while the rest of the body remains relatively still. For moderate intensity activities, the SLM, HLM, and HNLM all provided similar levels of classification accuracy overall (43.7–51.3%), with notable poor performances across all processing methods for stationary biking and walking while typing, activities where the wrists remain relatively still. However, the accuracy of estimating moderate intensity walking was excellent for the HLM and HNLM methods (92% and 90%, respectively), but poor for the SRF (9.8%). Finally, for vigorous intensity activities, each method had a high degree of accuracy ranging from 89% for the SLM to 99% for the SRF.

Discussion

Researchers are increasingly relying on wrist-worn accelerometer data to assess free-living PA as it may result in higher degrees of participant compliance. The evaluation of raw acceleration data (measured in gravitational units (g's) instead of counts) has also been of considerable

interest in recent years (Chen *et al* 2012, Troiano *et al* 2014). Reporting outcomes in g's offers potential advantages for the future since it facilitates comparisons and standardization across different monitoring devices. However, research is still in its infancy with this metric. Thus, understanding the strengths and weaknesses associated with different processing methods using raw data collected at the wrist is crucial. This is particularly true for AG accelerometers, which are the most widely used family of monitors in this type of research (Wijndaele *et al* 2015). To that end, this study examined the validity of four methods of processing raw, wrist-worn accelerometer data in relation to indirect calorimetry and the Compendium of PA in a sample of healthy adults.

Overall accuracy was relatively comparable among the 4 methods, with the SRF and HNLM providing similar and somewhat higher classification accuracy than the SLM and the HLM methods when averaged across activity intensities. Further, the results demonstrated that processing methods varied considerably in their ability to accurately classify activities by intensity as well as in their estimation of the MET values associated with different activities. As such, while none of the methods provided a high degree of accuracy across the board, some methods did better than others at capturing particular intensities and/or activities that may be of interest in future studies. Each of the methods relied on a different level of complexity in the model as well. As such, a summary of each model's distinctives and performance is offered below.

Hildebrand linear method

The Hildebrand equation represented the simplest method tested in the present study. A simple regression equation relying solely on ENMO, the Hildebrand equation performed relatively poorly on average, overestimating EE in sedentary and light activity and underestimating EE in moderate and vigorous activities. The overestimation of low activity levels is a result of the constant intercept in the linear equation that induces a floor value of approximately 2 METs. While this limits the ability for the HLM method to classify any activity as sedentary, it provides high accuracy in classifying all light standing activities correctly and was the most accurate measure for walking, which represents a large component of human movement. The variable accuracy at different intensities may also be related to the small, homogenous sample included in the Hildebrand study or the limited range of activity intensites used in its development. Classification agreement for this method was weaker than any other method, with the notable exception of vigorous intensity activity where this method was nearly perfect.

Hildebrand nonlinear method

The nonlinear, modified Hildebrand equation introduced an exponential term to model the relationship between acceleration and EE in a more complex fashion. This change resulted in substantive improvements over the original Hildebrand method, particularly at lower intensities. It exhibited more accurate estimates of mean EE and the lowest MAPEs of any of the processing methods for these lower intensities. However, the accuracy for the HNLM was relatively poor for tasks that disproportionately involved the wrist (e.g. throwing a ball) or tasks where the wrists were relatively motionless (e.g. standing, walking while typing), which we tested at light and moderate intensities. Improvement was also visible for vigorous activity, with the lowest MAPE of any processing method at this intensity. Improvements in classification accuracy compared to the original method were present, but heavily influenced by the new equation's ability to estimate sedentary behavior where its predecessor could not.

Table 2. Classification performance (minutes at each intensity) of wrist methods in comparison to Compendium, by intensity category and overall.

		Confusion matrix				Sensitivity				Specificity				Overall	
		Compendium													
		SED	LIGHT	MOD	VIG	SED	LIGHT	MOD	VIG	SED	LIGHT	MOD	VIG	Percent accuracy	Kappa CI ($\alpha = 0.05$)
HLM	SED	0	0	0	0	0%	67%	40%	98%	100%	37%	84%	99%	45.9%	(19%, 29%)
	LIGHT	199	172	123	0										
	MOD	9	83	82	2										
	VIG	0	1	0	100										
HNLM	SED	186	130	94	0	89%	21%	37%	98%	60%	90%	86%	99%	54.0%	(33%, 42%)
	LIGHT	15	54	35	0										
	MOD	7	71	76	2										
	VIG	0	1	0	100										
SLM	SED	81	34	20	0	39%	45%	44%	89%	90%	64%	74%	99%	48.9%	(24%, 35%)
	LIGHT	90	112	93	0										
	MOD	33	102	88	11										
	VIG	0	3	0	89										
SRF	SED	88	31	9	0	43%	48%	48%	99%	93%	60%	81%	99%	53.5%	(31%, 41%)
	LIGHT	105	121	95	0										
	MOD	11	95	97	1										
	VIG	0	4	0	99										

Note: percent accuracy = (correct estimates)/(total estimates).

Abbreviations: HLM—Hildebrand Linear Method; HNLM—Hildebrand Nonlinear Method; SLM—Staudenmayer Linear Model; SRF—Staudenmayer random forest; CI—confidence interval; SED—sedentary; MOD—moderate; VIG—vigorous.

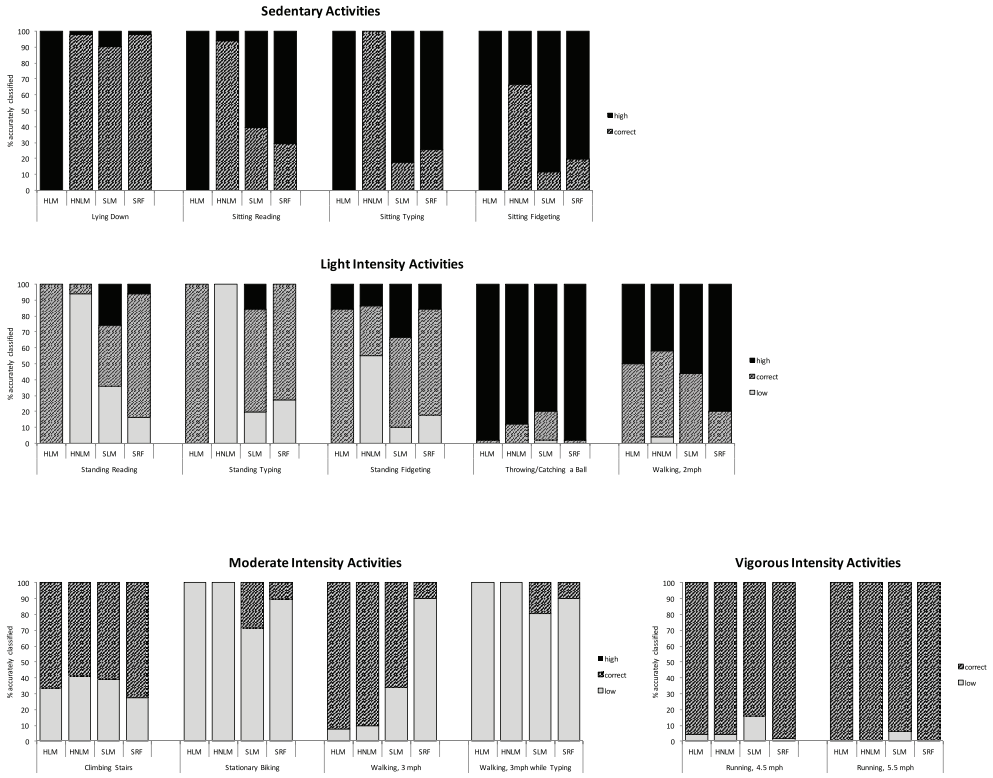

Figure 3. Agreement and accuracy for classifying activities by intensity for Hildebrand's linear method (HLM), the Hildebrand nonlinear method (HNLM), Staudenmayer's linear model (SLM), and Staudenmayer's random forest (SRF) method, for each of the 15 activities using the Compendium as the criterion.

Staudenmayer's linear method

Although still a regression equation, the SLM is more complex than either the HLM or the HNLM, both for its involvement of multiple terms, and for the significance of those terms. The model does not strictly relate acceleration to EE, but instead models the variability in signal (standard deviation of the vector magnitude) and inferred angular characteristics (mangle) to estimate EE. Descriptive analysis and MAPEs showed this method to be a poor predictor of EE at all intensities, with slight improvement at vigorous intensity. Bland–Altman plots for sedentary and light-intensity behaviors suggested systematic bias at low intensities. Classification agreement was generally low, with poor agreement overall and the lowest sensitivity to vigorous intensity activity of any of the wrist methods.

Staudenmayer's random forest method

The SRF method represented the most complex method tested in the present study, relying on a particular machine learning technique. The method showed similar overall descriptive characteristics to the SLM, with some improvements at lower intensities. Interestingly, it showed improvement at the individual level compared to the linear model with lower MAPEs, particularly at sedentary and light intensities. Whole-trial equivalence was nearly achieved,

but examination of intensity-specific equivalence showed the same pattern of overestimating at low intensities and underestimating at higher intensities as seen with the HLM and SLM methods, suggesting a poor ability for estimating EE. One strength of this approach is its effectiveness at detecting standing tasks, with better balance between sensitivity and specificity than the HLM. Classification agreement overall was stronger as compared with the linear models (HLM and SLM) and similar to the HNLM with more consistent sensitivity and specificity across intensities than the other methods.

The findings specific to the SLM and SRF methods are less promising than those reported by Staudenmayer and colleagues in their original validation paper (Staudenmayer *et al* 2015). However, the design of the two studies differed in ways that likely contribute to the discrepancies. The activities performed in the Staudenmayer study included the full range of intensities from sedentary to vigorous and incorporated a number of activities of daily living (e.g. folding laundry, gardening). However, there were no activities that involved standing relatively still, which is often misclassified as sedentary as opposed to light intensity activity. Nor did they include cycling, which is notably difficult to accurately capture with a wrist-worn accelerometer as the wrists are relatively stationary though EE is moderate.

This discrepancy between previous findings and the results from the present study highlights the need for further testing of all processing methods under free-living conditions to determine the accuracy across a range of activities carried out in normal daily life. Techniques for performing these types of comparisons have improved and analyses have used larger and more powerful data sets (Ellis *et al* 2016, Kerr *et al* 2016). However, the need for emphasis on intensity classification remains open.

The wrist is being increasing chosen over the hip for accelerometer placement in free-living studies based on evidence for superior compliance with study protocols in comparison to the hip in large-scale epidemiological studies (Freedson and John 2013, Troiano *et al* 2014); however, challenges with calibration and interpretation remain. A unique advantage of this study (and dataset) is that it was possible to directly compare the validity to values obtained from hip-worn accelerometer data from the same sample of participants (Ellingson *et al* 2016). There was similar variability across data processing methods, but several methods based on hip-worn accelerometer provided more favorable outcomes (Ellingson *et al* 2016). For example, we demonstrated that the Sojourns 3-axis method developed by Lyden *et al* (2014), which uses a single accelerometer worn at the hip had an overall classification accuracy of 56% and MAPE values for EE ranging from ~18% to 24% across intensities categories (Ellingson *et al* 2016). Further, the newly developed Sojourn Including Posture (SIP) method, which integrates data from a thigh-worn activPAL accelerometer into the data processing stream associated with Sojourns was better still, with an overall accuracy of 79% and MAPE values for EE ranging from ~15% to 20% (Ellingson *et al* 2016). It is important to point out that older linear regression methods for processing hip-worn accelerometer data using counts in conjunction with the more traditional cut-point method (Kim and Welk 2015) were less accurate than the wrist methods tested in this study with MAPE estimates ranging from ~24% to 100% (vertical axis) and ~21%–99% (vector magnitude) across intensities. Thus, the newly developed wrist methods (and those based on raw data) offer promise for continued refinement. The use of open-source methods and processing of raw data will ultimately help in improving standardization of methods.

Despite promising improvements in convergence, there were several activities for which all wrist methods were particularly poor. These included throwing and catching a ball as well as stationary biking and walking while typing. In each of these activities, the movement at the wrist was not likely representative of the body's movement as a whole, resulting in poor estimates of EE and reflecting an inherent limitation of wrist placement. These activities, in

particular, diminished the accuracy of all four methods at the moderate intensity classification. The same limitation would potentially extend to activities that were not included in the present protocol such as walking with hands in pockets, carrying heavy loads, or possibly even activities of daily living like folding laundry or loading a dishwasher. Likely, if more of these types of activities were assessed at light or vigorous intensities, the errors associated with those activity levels would similarly decline. Thus, the high errors observed with moderate intensity activities may be due to the specific activities selected rather than an inherent weakness of the methods to represent moderate intensity activities accurately. However, recent work by Ellis and colleagues demonstrates that pattern recognition approaches are improving and wrist-worn accelerometers are now able to differentiate between four activities including sitting, standing, walking/running, and riding in a vehicle with ~85% accuracy (Ellis *et al* 2016). It should be noted as well that a degree of this limitation may be present for any attachment site, including the hip. For example, postural tasks involving isometric contractions (including standing) require muscle energy without concomitant body accelerations, and many upper body resistance exercises may be misclassified as sedentary when the sensor is placed at the hip. Thus, as instruments converge and approach their current ceiling for utility, focus should shift to making that ceiling higher, perhaps by introducing related sensors (e.g. gyroscopes, heart rate sensors) into existing instruments and applying similar pattern recognition techniques.

There are several limitations inherent in our study. Participants in our sample were primarily young adults with normal BMI. As such these processing methods should undergo additional testing in both younger and older individuals as well as those in the overweight and obese categories to determine its validity in these populations. Further, BMR was not objectively measured; rather it was estimated using the Schofield equation (Schofield 1985). However, this is a relatively common and accepted practice in validation studies since it tends to reduce some individual error (Carter *et al* 2008, Westerterp 2013, Kim and Welk 2015). Additionally, accelerometers were worn only on the right wrist. Thus, although there is evidence that results do not differ substantively based on handedness (Troiano *et al* 2014, Montoye *et al* 2015), the results may not reflect those from studies where participants wore accelerometers on their non-dominant wrist, as is done in the NHANES. Future validation studies should include accelerometers worn on both wrists to assess the relative differences between dominant and non-dominant and to make recommendations regarding the best placement for research studies (Troiano *et al* 2014). Further, the mean error rates at each intensity are highly dependent on the tasks chosen. Thus, these error rates are not necessarily representative of all activities at each intensity, but provide a range of types and intensities to better assess the strengths and weaknesses of each methodology. Lastly, this study did not include a free-living component. However, the 15 activities included in the protocol ranged from sedentary to vigorous in order to capture a range of potential activities individuals may engage in during normal daily life, as well as several light and moderate intensity activities that were expected to show a discrepancy between wrist motions and EE. Nonetheless, additional testing of these methods under free-living conditions including household activities will be an important next step.

In summary, we present the validity of four processing methods for handling raw data generated by wrist-worn AG accelerometers. Our results demonstrated that across activities of varying intensities, none of the methods had a high level of agreement with criterion measures. Thus results from wrist-worn accelerometry should continue to be interpreted with caution. Further refinements of these methods and/or the development of new methods to process wrist-worn accelerometer data are warranted to improve our understanding regarding the impact of PA and sedentary behaviors on health.

Source of funding

Work by the Iowa State University authors was funded by the National Cancer Institute under contract number 6053-S03 issued to Westat.

Conflicts of interest

No authors have any conflicts of interest to report. The results of this study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation.

References

Ainsworth B E *et al* 2011 2011 compendium of physical activities: a second update of codes and MET values *Med. Sci. Sports Exerc.* **43** 1575–81

Anon 2016 *R: A Language and Environment for Statistical Computing* Available at: https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf (Accessed: 29 April 2016)

Bassett D R, Rowlands A and Trost S G 2012 Calibration and validation of wearable monitors *Med. Sci. Sports Exerc.* **44** S32–8

Carter J *et al* 2008 An investigation of a novel three-dimensional activity monitor to predict free-living energy expenditure *J. Sports Sci.* **26** 553–61

Chen K Y *et al* 2012 Redefining the roles of sensors in objective physical activity monitoring *Med. Sci. Sports Exerc.* **44** S13–23

Da Silva I C *et al* 2014 Physical activity levels in three Brazilian birth cohorts as assessed with raw triaxial wrist accelerometry *Int. J. Epidemiol.* **43** 1959–68

Ellingson L D *et al* 2016 Validity of an integrative method for processing physical activity data *Med. Sci. Sports Exerc.* **48** 1629–38

Ellis K *et al* 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

Ellis K *et al* 2016 Hip and wrist accelerometer algorithms for free-living behavior classification *Med. Sci. Sports Exerc.* **48** 933–40

Freedson P S and John D 2013 Comment on 'estimating activity and sedentary behavior from an accelerometer on the hip and wrist' *Med. Sci. Sports Exerc.* **45** 962–3

Hildebrand M *et al* 2014 Age group comparability of raw accelerometer output from wrist- and hip-worn monitors *Med. Sci. Sports Exerc.* **46** 1816–24

Hodges L D, BrodieD A and Bromley P D 2005 Validity and reliability of selected commercially available metabolic analyzer systems *Scand. J. Med. Sci. Sports* **15** 271–9

Kerr J *et al* 2016 Objective assessment of physical activity: classifiers for public health *Med. Sci. Sports Exerc.* **48** 951–7

Kim Y and Welk G J 2015 Criterion validity of competing accelerometry-based activity monitoring devices *Med. Sci. Sports Exerc.* **47** 2456–63

Liaw A and Wiener M 2002 Classification and regression by randomForest *R News* 2/3 pp 18–22

Lyden K *et al* 2014 A method to estimate free-living active and sedentary behavior from an accelerometer *Med. Sci. Sports Exerc.* **46** 386–97

Montoye A H K *et al* 2015 Energy expenditure prediction using raw accelerometer data in simulated free living *Med. Sci. Sports Exerc.* **47** 1735–46

Salier Eriksson J, Rosdahl H and Schantz P 2012 Validity of the Oxycon mobile metabolic system under field measuring conditions *Eur. J. Appl. Physiol.* **112** 345–55

Schofield W N 1985 Predicting basal metabolic rate, new standards and review of previous work *Hum. Nutr. Clin. Nutr.* **39** 5–41

Staudenmayer J *et al* 2015 Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements *J. Appl. Physiol.* **119** 396–403

Troiano R P *et al* 2014 Evolution of accelerometer methods for physical activity research *Br. J. Sports Med.* **48** 1019–23

UKBiobank 2016 Available at: www.ukbiobank.ac.uk/physical-activity-monitor/ (Accessed: 25 March 2016)

Van Hees V T *et al* 2014 Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents *J. Appl. Physiol.* **117** 738–44

Westerterp K R 2013 Physical activity and physical activity induced energy expenditure in humans: measurement, determinants, and effects *Front. Physiol.* **4** Available at: www.ncbi.nlm.nih.gov/pmc/articles/PMC3636460/ (Accessed: 5 October 2015)

Wijndaele K *et al* 2015 Utilization and harmonization of adult accelerometry data: review and expert consensus *Med. Sci. Sports Exerc.* **47** 2129–39