Article:

O'Driscoll, R, Turicchi, J orcid.org/0000-0003-1174-813X, Hopkins, M orcid.org/0000-0002-7655-0215 et al. (3 more authors) (2020) Improving energy expenditure estimates from wearable devices: A machine learning approach. Journal of Sports Sciences, 38 (13). pp. 1496-1505. ISSN 0264-0414

https://doi.org/10.1080/02640414.2020.1746088

Improving Energy Expenditure Estimates From Wearable Devices: A Machine Learning Approach

O'Driscoll Ruairi[1], Turicchi Jake[1], Hopkins Mark[2], Horgan Graham. W.[3], Finlayson Graham[1], Stubbs James. R.[1].

[1]*Appetite Control and Energy Balance Group, School of Psychology, University of Leeds, Leeds, U.K*

[2] *School of Food Science and Nutrition, Faculty of Mathematics and Physical Sciences, University of Leeds, Leeds, U.K.*

[3] *Biomathematics & Statistics Scotland, Aberdeen, U.K.*

Email addresses:

O'Driscoll Ruairi (psrod@leeds.ac.uk), Turicchi Jake (psjt@leeds.ac.uk) , Hopkins Mark (M.Hopkins@leeds.ac.uk), Horgan Graham. W. (graham.horgan@bioss.ac.uk), Finlayson Graham (g.s.finlayson@leeds.ac.uk) , Stubbs James. R. (r.j.stubbs@leeds.ac.uk).

Keywords:

Machine learning,

Heart rate,

Energy Expenditure,

Accelerometer

Abstract

A means of quantifying continuous, free-living energy expenditure (EE) would advance the study of bioenergetics. The aim of this study was to apply a non-linear, machine learning algorithm (random forest) to predict minute level EE for a range of activities using acceleration, physiological signals (e.g. heart rate, body temperature, galvanic skin response), and participant characteristics (e.g. sex, age, height, weight, body composition) collected from wearable devices (Fitbit charge 2, Polar H7, SenseWear Armband Mini and Actigraph GT3-x) as potential inputs. By utilising a leave-one-out cross-validation approach in 59 subjects, we investigated the predictive accuracy in sedentary, ambulatory, household, and cycling activities compared to indirect calorimetry (Vyntus CPX). Over all activities, correlations of at least $r=0.85$ were achieved by the models. Root mean squared error ranged from 1-1.37 METs and all overall models were statistically equivalent to the criterion measure. Significantly lower error was observed for Actigraph and Sensewear models, when compared to the manufacturer provided estimates of the Sensewear Armband ($p<0.05$). A high degree of accuracy in EE estimation was achieved by applying non-linear models to wearable devices which may offer a means to capture the energy cost of free-living activities.

Background

The measurement of energy expenditure (EE) is critical to understand human energy requirements in health and disease, and how components of energy balance contribute to over and under nutrition. Quantifying energy balance in free-living individuals requires the precise and accurate estimation of at least two of the three components of energy balance; energy intake (EI), EE and changes in energy stored (ES). Currently, quantifying true patterns of EI and EE in the free-living environment is constrained by methodological and practical limitations. Objective measures of EI can be derived from EE and changes in ES (e.g. the intake-balance method [1]), but the use of doubly labelled water (DLW) to estimate EE is costly and fails to capture daily variation in EE, which limits its widespread adoption [2].

Activity monitors have long been recognised as a potential means to estimate EE [3], but the accuracy of current devices in estimating the energy cost of a wide range of activities and intensities is limited [4]. Accelerometery is routinely used to quantify bodily movement and to predict EE using linear models [5,6]. However, the relationship between EE and acceleration is variable between activities [7], and accelerometery alone has limited sensitivity to capture the additional energy demands of activities that do not alter the velocity of movement (e.g. load carrying or incline walking) [8]. While estimates of EE from devices with HR sensing technology are likely to have reduced error relative to devices based on accelerometery alone [4], the relationship between accelerometery, HR and EE exhibits linearity only within specific activity types. Combined linear models that estimate EE will therefore not generalise across the range of human activities [9]. In some cases, this may explain the demonstrably poor manufacturer provided EE estimates obtained from many current and past activity monitors [10].

Complex, non-linear machine learning algorithms applied to research-grade accelerometers have shown remarkable accuracy in estimating EE using tree-based methods

[11] and artificial neural networks [9,12] and they may offer a means to overcome the limited accuracy of current devices. However, whether machine learning can be used to improve the estimation of EE using the sensor data obtained from commercial devices has yet to be examined.

Machine learning methods demonstrate clear potential to estimate EE using data obtained from wearable sensors. This study aims to explore the potential for non-linear, machine learning regression models utilising subject characteristics, movement and physiological variables to estimate EE in a range of activities.

Methods

Participants

A sample of 59 participants were included (Female=41, Age = 44.4 ± 14.1 years, Weight = 75.7 ± 13.6 kg, BMI = 26.9 ± 4.7 kg/m^2, FM= 24.8 ± 10.73 kg, FFM = 49.8 ± 8.9, FM (%) = 32.5 ± 10.3%, FFM (%) = 67.5 ± 10.3%, RMR = 1581.8 ± 280.4 kcal/d). Participants were primarily from the Leeds centre of the NoHoW trial (n = 44), ISRCTN registry (ISRCTN88405328), an additional 15 participants were recruited from the University of Leeds and surrounding areas. Participants were excluded from the study for the following reasons: pregnancy, medications altering metabolic rate, cardiovascular, metabolic, renal disorders, illness or injury that provide an increased risk of medical events during PA [13]. Ethical approval was granted by The University of Leeds, School of Psychology ethics committee (PSC-407, 18/08/2018).

Physical measurements

Measurements were conducted at an exercise laboratory at The Human Appetite Research Unit, University of Leeds. Participants arrived at the university between 06:00 am and 09:30 am, having refrained from the intake of food, caffeine and exercise for 12 hours prior to the measurements. Systolic and diastolic blood pressure (BP) and resting heart rate (HR) (Microlife BP A2 Basic, Gentle Technology, Microlife, Clearwater, FL, USA, Inc.) were measured at rest and in the sitting position. Height (±0.1cm) was measured barefoot, using a Seca 704s instrument (SECA, Germany). Fat mass (FM) and fat-free mass (FFM) were estimated using air displacement plethysmography (BodPod, Life Measurement, Inc.; USA) and the Siri equation [14]. Body weight (± 0.1kg) was also obtained from the BodPod scales in light clothing.

Table 1 – insert here

Physical activity devices

Participants wore a number of physical activity devices during the study and all devices were initialised in accordance with manufacturer's instructions. The Polar m400 HR Monitor Watch and a Polar H7 chest strap (Polar Electro, Kempele, Finland) were used to measure HR. The Polar H7 facilitates approximately 200 hours of continuous transmission. In this study data were extracted at the second level and averaged to the minute-level. Participants also wore a Fitbit Charge 2 (FC2) (Fitbit Inc, San Francisco, CA, USA), a wrist-worn activity monitor, which incorporates a tri-axial accelerometer. The FC2 also estimates HR through a patented technology called 'PurePulse', which uses light-emitting diodes to measure changes to blood volume [15]. An Actigraph GT3-x accelerometer (AG) was placed on the non-dominant wrist which measured acceleration along vertical, horizontal and perpendicular axes at a sample rate of 30Hz. Participants also wore the SenseWear Armband Mini (SWA) (BodyMedia Inc., Pittsburgh, PA) on the non-dominant upper arm. The SWA collected tri-axial accelerometer data and as well as data from heat-related sensors measuring heat flux, skin temperature, near body ambient temperature and galvanic skin response.

Energy expenditure measurement

Resting metabolic rate (RMR) was measured using an indirect calorimetry system fitted with a ventilated hood (GEM, Nutren Technology Ltd; UK). Participants lay in the supine position for 30 minutes, whilst VO_2 and VCO_2 were continually measured. An RMR estimate was derived from 5 minutes of steady state data, as described previously [16]. Briefly, after discarding the first 5 minutes, the VO_2 and VCO_2 measurements in the 5-minute period with the lowest coefficient of variation during the overall measurement period are used to estimate RMR. In the absence of an RMR measurement (n=2), a body mass index (BMI) specific RMR algorithm was used [17]. A stationary metabolic cart (Vyntus CPX, Jaeger-CareFusion, UK) was used as the criterion measure of EE in the physical activity protocol. Breath by breath data were aggregated to minute level to estimate EE (kcal/min^{-1}). The Vyntus CPX is highly valid and reliable [18,19] and served as a criterion comparison for the developed models. The unit was calibrated prior to each lab visit in accordance with manufacturer's instructions. Data were aggerated to the minute level and EE (kcal/min^{-1}) values were calculated from VO_2 and VCO_2 data assuming a minimal contribution of protein oxidation [20]. We expressed minute level EE as a multiple of each participant's RMR, to derive metabolic equivalents (METs), which was the outcome variable.

Physical activity protocol

Participants undertook a structured protocol consisting of 10 activities, which were performed at a consistent intensity of 5 minutes each. The activities were performed in a set order and included: sitting, standing, treadmill walking (4 km/h), incline treadmill walking (4 km/h, 5% incline), jogging (6-8 km/h, 5% incline), incline jogging (6-8 km/h, 5% incline). Next, after a 3-minute resting period, participants transitioned to a cycle ergometer for low-intensity (30 watts), and moderate intensity cycling (60 watts). After another period of recovery, participants performed a folding task and lastly a sweeping task. The physical activity protocol was performed by all participants, however the jogging task (n=49), the jogging 5% incline (n=30) and the moderate cycling tasks (n=58) were not performed by all participants, due to variation in physical fitness.

Data processing and model development

All data sources were aggregated to the minute-level and were matched by time for each participant. The first minute of data was removed leaving minutes 2-5 for inclusion in model development [11,21]. The models developed in this study were trained using complete minute level data only, so if any data points were missing from the sensors or subject characteristics, that single minute was not included in the analysis. In the present study, we developed distinct predictive models for each device (FC2, AG and SWA) and the specific predictor variables used in each of these is described in table 1. The algorithm used in the present study was a random forest regressor [22]. Random forests are an ensemble method which aggregate the output of numerous decision trees to produce a continuous output. In the random forest algorithm, trees are trained on a random sample of the available predictor variables, which reduces the chances overfitting the training data [23]. For all the random forest models, the number of variables randomly sampled at each split was set 1/3 of the number of predictor variables in the model. For each of the developed models, 1000 trees were grown, and minimum size of terminal nodes was set to 5. All model development and training was conducted with the "randomForest" package [24] in R. Model parameters were established in preliminary tuning experiments and were standardised to allow comparability between each of the models.

Statistical analysis

Two validation approaches were used in this study. A 'holdout' approach was used in which all available data are split into training and testing, at a ratio of 80:20. Secondly, Leave-One-Out Cross-Validation (LOOCV), in which models are trained on all participants' data with the exception of one participant, which serves as the testing dataset. This process is repeated until all participants data has been used to test the algorithm.

A range of statistical tests were employed to investigate the accuracy of model estimates, in line with previous validation research [25]: Pearson's correlation coefficients, root mean squared error (RMSE) and mean absolute percentage error (MAPE), calculated with the R package 'metrics' [26]. Equivalence tests were used to determine whether the models were statistically equivalent to the criterion measured METs, to be considered equivalent, the 90% confidence interval of the estimate must fall within ± 10% of the criterion mean [27]. Repeated measures analysis of variance (ANOVA) tests were employed to test for differences in MAPE calculated for each of the models, and the SWA for each subject's activity modality. We investigated differences between specific models with pairwise t-tests conducted with a Holm–Bonferroni false error rate correction. All data are reported as means and standard deviations (SD) unless otherwise stated. In order to estimate the precision of estimates in this study, standard errors for the overall RMSE of each model have been computed at the participant level and are presented in supplementary table 1.

All analyses and data processing were conducted R version 3.5.1 and RStudio Version 1.1.447 [28], using a p-value of < 0.05 to determine statistical significance.

Results

Predictive accuracy of models

The performance of the holdout validation was typically superior to the LOOCV method, as measured by MAPE. The correlation of all models exceeded r values of 0.94 and the results of the models using the holdout approach are shown in supplementary table 2.

Models $FBRF_1$ and $FBRF_2$ demonstrated the greatest MAPE and highest RMSE using this approach.

Table 2 – insert here

The performance of the models without body composition data ($AGRF_2$, $SWRF_2$, $FBRF_1$) using a LOOCV validation approach are shown in the form of scatterplots in figure 1. The accuracy statistics from the LOOCV validation and the results of the equivalence tests are presented in table 2. Data loss occurred for 2 participant's FC2 data, one participant's AG data and one participant's polar HR data. All models were validated on at least 2000 minutes and 55 participants and individual level data is presented in supplementary table 3. The SWA was not statistically equivalent to the criterion measure, in contrast to the random forest models, which were all statistically equivalent to the criterion measure. The SWA also had the highest RMSE of 1.8 METs compared to the AGRF and SWRF models which ranged between 1-1.24 METs and FBRF models, which had RMSE values of 1.37 METs or less.

Figure 1 – insert here

Table 2 – insert here

The results of the ANOVA demonstrated a significant F statistic of 41.79 ($p= 7.26^{-49}$) for between model differences, indicating that differences existed between the MAPE values for each model's METs estimates relative the criterion METs. Pairwise t-tests demonstrated that the MAPE for the SWA estimates were significantly higher than all random forest models ($p<0.05$), except for FBRF models. Model $AGRF_1$ had MAPE values significantly higher than $AGRF_2$ ($p=1.16^{-10}$) and $AGRF_3$ ($p=2.03^{-08}$). $SWRF_1$ was significantly higher than $SWRF_2$ ($p=7.19^{-11}$) and $SWRF_3$ ($p=4.14^{-09}$).

The introduction of body composition did not result in significantly different MAPE models developed on AG sensor outputs, however, $FBRF_1$ had a significantly lower MAPE than the body composition model, $FBRF_2$ ($p=0.007$) and $SWRF_2$ was significantly lower than

$SWRF_3$ (p=0.021), indicating a less accurate model performance with the addition of body composition. All SWRF and AGRF models had significantly lower MAPE values than $FBRF_1$ and $FBRF_2$ (p<0.01).

Activity specific accuracy

Activity specific accuracy statistics calculated using LOOCV are presented in table 3. The accuracy of the FBRF models was poorest in sedentary tasks, where MAPE values of 47.87 and 52.30 were observed for $FBRF_1$ and $FBRF_2$, respectively during the standing task. Both $FBRF_1$ and $FBRF_2$ were statistically equivalent during the jogging task. Models $AGRF_2$ and $AGRF_3$ were statistically equivalent in 5 of 10 tasks, namely: standing, incline walking, jogging, low intensity cycling and folding, and the $AGRF_1$ model was equivalent in all of the aforementioned tasks, except from incline walking. The MAPE values ranged from 13.01 ($AGRF_3$, walk incline) to 29.33 ($AGRF_1$, sweeping). Models developed on SWA data were statistically equivalent for walking ($SWRF_2$ and $SWRF_3$ only), walking incline, jogging and low intensity cycling, and the $SWRF_2$ was equivalent in sitting. $SWRF_2$ and $SWRF_3$ demonstrated the poorest accuracy in the household tasks with overestimates in models and MAPE values ranging from 24.11 to 33.41.

Table 3 – insert here

Model characteristics

Using the feature set which included body composition data for each device ($AGRF_3$, $FBRF_2$, $SWRF_3$), we computed the relative importance of each predictive variables [22]. The variable in the plots represents the percentage increase to the mean squared error following the permutation (random shuffling) of each variable. Permutation in this manner breaks the association between the predictive and outcome variable relative to the original model and therefore facilitates estimates of the importance of this variable to overall accuracy of the original model. Outlined in figure 2 for the FB, acceleration (i.e. steps) and HR normalised to the sitting HR were the most important variables in the models, with age, kilograms of FFM, height and FM following after. In the $SWRF_3$ and $AGRF_3$ models (figure 2), HR was associated with the greatest increase in mean squared error.

<u>Figure 2 – insert here</u>

Discussion

This is the first study to demonstrate that sensor data obtained from commercial wearable devices (i.e. FC2) can be used to estimate EE with a high degree of validity in a diverse range of activities. Commercial activity monitors offer a number of benefits over research-grade devices, including their economic viability, participant acceptance and cloud storage capabilities [29] and our findings highlight the potential for these inexpensive tools to more accurately quantify EE. We show that accelerometer data collected from research-grade devices on the wrist or arm, can be used to predict EE with a high degree of accuracy in a diverse population.

The results of the present study are comparable to the accuracy reported in previous studies, with overall RMSE reaching 1 MET for the most accurate model ($SWRF_2$) and 1.37 METs for the least accurate model ($FBRF_2$). Using Actigraph accelerometer data, Ellis et al. trained random forest models and reported a RMSE of 1 METs for bout predictions from a hip worn accelerometer, 1.09 from a wrist accelerometer [11]. Staudenmayer et al. developed artificial neural networks and report a RMSE of 1.22 METs [9]. Montoye et al., showed that artificial neural networks trained on wrist accelerometer data were more accurate than corresponding linear models, with RMSE values between 1.26–1.32 METs [12]. Importantly however, this is the first study to apply machine learning to the sensor data obtained from a commercially available tracking device (e.g. FC2) in order to improve the accuracy of EE estimates. In the present study, the RMSE for FBRF models approached the accuracy of the models developed using research grade accelerometers, indicating that models developed using minute-level data from wearable sensors in combination subject characteristics can be modelled to accurately predict EE. That said, the activities performed in our protocol are generally less diverse than some of the aforementioned studies.

The models with more accelerometer variables as input features (AGRF and SWRF) led to the greatest predictive accuracy, indicating the importance of tri-axial accelerometery. The variable importance plots show that HR was the most important determinant of error reduction for the $SWRF_3$ and $AGRF_3$ models, which reflects the established relationship between HR and VO_2 [30,31]. In contrast, the $FBRF_2$ model was less influenced by this variable and we postulate that this may be related to the sensor used to collect HR estimates. The polar HR strap, which was used in two AGRF and SWRF models shows near perfect agreement with electrocardiogram criterion measures [32]. Conversely, photoplethysmography based HR sensors may produce ‘spurious’ HR measurements [10], which increases noise in the training and testing data sets and had a detrimental effect on the predictive accuracy.

We normalised HR to each participant’s sitting HR and used this as a predictor in the developed models. This was motivated by the established relationship between the sitting HR and the flex point [33]. Given the important predictive role this variable has on the models, the use of HR above sitting appears to offer a means of capturing some of the individual variability in the relationship between VO_2 and HR without the need for individual calibration, as in previous approaches [30]. Despite the importance of HR in achieving the highest predictive accuracy, we show that AG and SWA models developed without external HR data can still produce highly accurate estimates of EE, surpassing the manufacturer estimates of the SWA and FBRF models. Considering the potential burden of additional wearable devices (i.e. chest HR straps), this finding has implications for population research in which the use of a single device is of considerable appeal.

The addition of body composition resulted in a significantly higher MAPE in the case of SWRF models and this may be explained by the specifics of the random forest algorithm. Each of the trees grown in the random forest regressor samples approximately 1/3 of the predictor variables and a bootstrapped sample of the training data, which serves to decorrelate the trees and limits the likelihood of overfitting. Introducing body composition could result in a situation in which splits are less likely to include the most relevant predictor variables such as HR, which theoretically could have a detrimental effect on the global model [34]. Furthermore, body composition (FFM) is highly correlated with RMR [35], it is likely that any variance attributable to body composition is already accounted for by normalising our predictions relative to RMR. Thus, FFM may have a greater predictive ability if absolute caloric expenditure, rather than METs, was the outcome measure. Regardless, the accurate and precise measurement of body composition is time consuming, costly and requires experimental expertise. In this sense, the finding that body composition is not critical to the predictive accuracy of the random forests has positive implications for the utility of the models in large studies.

We included ambulatory, resting, cycling and housework tasks, which are challenging to assess using wearable devices owing to the differing accelerometer patterns produced by each activity [36]. It is notable that the AGRF and SWRF models were statistically equivalent with the criterion measure (indirect calorimetry) in a many activity modalities. This demonstrates the potential of a single non-linear model to accurately estimate the bioenergetic demands of common activity types and overcome the limitations of traditional linear approaches, which have a tendency to generalise poorly across the spectrum of human activities [36–38]. However, equivalence was not achieved in all activities; activity specific prediction may be enhanced by generating larger training data or combining activity classification with regression models.

The SWA is considered one of the more valid wearable devices for estimating TDEE [39–41], and a recent meta-analysis from our group showed that this monitor was one of the only wrist or arm-worn activity monitoring devices not to systematically over or underestimate EE overall relative to DLW [4]. Thus, we compared the SWA METs to models developed in this study to facilitate the interpretation of the developed models. All models had lower MAPE, RMSE and higher correlations to the criterion measure than the SWA. These data therefore suggest that machine learning can be applied to sensor data of commercially available devices to surpass the accuracy of widely used research-grade devices. The SWA provides sufficiently accurate estimates of EE that it can be used with measures of body weight/composition to estimate true EI [42]. Unfortunately, studies such as this are limited in their duration owing to the data storage capabilities and battery life of the SWA. Our results indicate that it may be possible to replace the SWA with machine learning models applied to commercial devices, i.e. FC2. Given that these data are accessible continually from the Fitbit API, this could offer an opportunity for a new generation of quantitative, long term energy balance research. Mathematical models developed to predict EI from body weight have been proposed and demonstrate a high degree of accuracy compared to EI calculated through DLW and DEXA [43,44]. In addition to the cost associated with these techniques, a recognised limitation of this model is the lack of a continuous EE estimate. Refining estimates of EE would improve the accuracy of such models and provide important data on day-to day variability in physical activity behaviours and associated EE currently not measurable by the DLW method [45].

A benefit of the present study is the expression of EE in METs relative to each participant's measured RMR. The assumption that 1 MET is equivalent to 3.5 ml O_2/kg$^-$

^{1}/min^{-1} can result in substantial bias, depending on the age and body composition of the subject in question [46]. Secondly, we report the results of two validation methods, a holdout approach and LOOCV approach. We envisage different experimental protocols in which participant's data may be available for a calibration procedure prior to beginning an observation period, as is practiced in the historical 'flex' method [30]. In this situation, the holdout method may be more reflective of the potential accuracy. It is more probable that individual calibration would not be possible, and the models would be applied to unseen participants, in this case a LOOCV is more appropriate. Thirdly, we computed accuracy statistics from all available minutes in the dataset, rather than aggregating them to 'activity bouts'. Indeed averaging in this manner has the potential to smooth errors and result in an artificially low average error. Human activity is performed for different durations and it is therefore valuable to determine accuracy at the minute-level.

Several limitations of this study should be acknowledged, firstly, approximately 70% of our sample were female. Secondly, the confinement to a laboratory and the utilisation of only steady state activity minutes may limit the ecological validity of this study and it therefore remains uncertain how well these models will perform in a less controlled environment with different activity types. It will be important to validate the models against whole-room calorimetry and the DLW method. Thirdly, we made no attempt to impute missing data in this study; devices will be removed or may fail in real life situations this is likely to create missing data. Considering these limitations, it is of great importance to continuously test and refine the presented models using data collected from different sedentary and active behaviours, participants, devices and durations. This is particularly important for commercial devices as updated and/or new models regularly come to market; nevertheless, the utilisation of three devices in this study indicates that the modelling approach taken would be applicable to newer devices.

Conclusion

This study demonstrates the potential for machine learning models developed using minute-level data from wearable sensors in combination subject characteristics to be modelled to predict EE with minimal bias. Further, machine learning models using outputs from a commercial activity monitor achieve greater predictive accuracy than the SWA armband. This methodology opens the possibility for quantitative energy balance research with affordable, unobtrusive wearable sensors.

References

1. Racette SB, Das SK, Bhapkar M, Hadley EC, Roberts SB, Ravussin E, et al. Approaches for quantifying energy intake and %calorie restriction during calorie restriction interventions in humans: the multicenter CALERIE study. AJP Endocrinol Metab [Internet]. 2012;302:E441–8. Available from: http://ajpendo.physiology.org/cgi/doi/10.1152/ajpendo.00290.2011

2. Black AE, Cole TJ. Within- and between-subject variation in energy expenditure measured by the doubly-labelled water technique: Implications for validating reported dietary energy intake. Eur J Clin Nutr [Internet]. 2000;54:386–94. Available from: http://www.nature.com/doifinder/10.1038/sj.ejcn.1600970

3. Jakicic JM, Marcus M, Gallagher KI, Randall C, Thomas E, Goss FL, et al. Evaluation of the SenseWear Pro Armband to Assess Energy Expenditure during Exercise. Med Sci Sport Exerc [Internet]. United States; 2004;36:897–904. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00005768-200405000-00024

4. O'Driscoll R, Turicchi J, Beaulieu K, Scott S, Matu J, Deighton K, et al. How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. Br J Sports Med [Internet]. BMJ Publishing Group Ltd and British Association of Sport and Exercise Medicine; 2018 [cited 2018 Oct 1];77:bjsports-2018-099643. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30194221

5. Freedson PS, Melanson E, Sirard J. Calibration of the Computer Science and Applications, Inc. accelerometer. Med Sci Sports Exerc. United States; 1998;30:777–81.

6. Crouter SE, Bassett DRJ. A new 2-regression model for the Actical accelerometer. Br J Sports Med. England; 2008;42:217–24.

7. Bonomi AG, Plasqui G, Goris AHC, Westerterp KR. Improving assessment of daily

energy expenditure by identifying types of physical activity with a single accelerometer. J Appl Physiol [Internet]. 2009;107:655–61. Available from: http://www.physiology.org/doi/10.1152/japplphysiol.00150.2009

8. Hills AP, Mokhtar N, Byrne NM. Assessment of Physical Activity and Energy Expenditure: An Overview of Objective Measures. Front Nutr [Internet]. 2014;1:1–16. Available from: http://journal.frontiersin.org/article/10.3389/fnut.2014.00005/abstract

9. Staudenmayer J, Pober D, Crouter S, Bassett D, Freedson P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. J Appl Physiol [Internet]. United States: American Physiological Society; 2009 [cited 2019 Feb 5];107:1300–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19644028

10. Reddy RK, Pooni R, Zaharieva DP, Senf B, El Youssef J, Dassau E, et al. Accuracy of Wrist-Worn Activity Monitors During Common Daily Physical Activities and Types of Structured Exercise: Evaluation Study. JMIR mHealth uHealth [Internet]. JMIR mHealth and uHealth; 2018 [cited 2018 Dec 16];6:e10338. Available from: https://mhealth.jmir.org/2018/12/e10338/

11. Ellis K, Kerr J, Godbole S, Lanckriet G, Wing D, Marshall S. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. Physiol Meas [Internet]. England; 2014 [cited 2019 Feb 5];35:2191–203. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25340969

12. Montoye AHKK, Begum M, Henning Z, Pfeiffer KA. Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data. Physiol Meas. IOP Publishing; 2017;38:343.

13. ACSM. Exercise Preparticipation Health Screen Recommendations. 2018 [cited 2018 Feb 20]; Available from: http://www.acsm.org/docs/default-source/publications/acsm-101-

prescreeninginfographiccolorlegal-2015-12-15-v02.pdf?sfvrsn=2

14. William E. Siri. BODY COMPOSITION FROM FLUID SPACES AND DENSITY: , ANALYSIS OF METHODS. Adv Biol Med Phy. 1956;

15. Benedetto S, Caldato C, Bazzan E, Greenwood DC, Pensabene V, Actis P. Assessment of the fitbit charge 2 for monitoring heart rate. PLoS One [Internet]. 2018;13:e0192691. Available from: http://dx.plos.org/10.1371/journal.pone.0192691

16. Sanchez-Delgado G, Alcantara JMA, Ortiz-Alvarez L, Xu H, Martinez-Tellez B, Labayen I, et al. Reliability of resting metabolic rate measurements in young adults: Impact of methods for data analysis. Clin Nutr [Internet]. 2018 [cited 2019 Apr 24];37:1618–24. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0261561417302674

17. Müller MJ, Bosy-Westphal A, Klaus S, Kreymann G, Lührmann PM, Neuhäuser-Berthold M, et al. World Health Organization equations have shortcomings for predicting resting energy expenditure in persons from a modern, affluent population: generation of a new reference standard from a retrospective analysis of a German database of resting energy expe. Am J Clin Nutr [Internet]. 2004;80:1379–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15531690

18. Perez-Suarez I, Martin-Rincon M, Gonzalez-Henriquez JJ, Fezzardi C, Perez-Regalado S, Galvan-Alvarez V, et al. Accuracy and Precision of the COSMED K5 Portable Analyser. Front Physiol [Internet]. Frontiers; 2018 [cited 2019 Apr 1];9:1764. Available from: https://www.frontiersin.org/article/10.3389/fphys.2018.01764/full

19. Groepenhoff H, de Jeu RC, Schot R. Vyntus CPX compared to Oxycon pro shows equal gas-exchange and ventilation during exercise. Respir Funct Technol [Internet]. European Respiratory Society; 2017 [cited 2019 Apr 24]. p. PA3002. Available from: http://erj.ersjournals.com/lookup/doi/10.1183/1393003.congress-2017.PA3002

20. Péronnet F, Massicotte D. Table of nonprotein respiratory quotient: an update. Can J

Sport Sci [Internet]. 1991 [cited 2019 Jun 11];16:23–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1645211

21. Staudenmayer J, Pober D, Crouter S, Bassett D, Freedson P. An artificial neural network to estimate physical activity energy expenditure and identify physical activity type from an accelerometer. J Appl Physiol. 2009;107:1300–7.

22. Breiman L. Random Forests. Mach Learn [Internet]. Kluwer Academic Publishers; 2001 [cited 2019 Feb 9];45:5–32. Available from: http://link.springer.com/10.1023/A:1010933404324

23. Tibshirani TH and R. Introduction to Statistical Learning. Japan Med. Assoc. J. 2011.

24. Liaw A, Wiener M, Andy Liaw M. Package: "randomForest". 2018 [cited 2019 May 13]; Available from: https://www.stat.berkeley.edu/~breiman/RandomForests/

25. Bai Y, Hibbing P, Mantis C, Welk GJ. Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR. J Sports Sci [Internet]. Routledge; 2018;36:1734–41. Available from: https://doi.org/10.1080/02640414.2017.1412235

26. Hamner B, Frasco M, Ledell E. Package "Metrics" Title Evaluation Metrics for Machine Learning [Internet]. 2018. Available from: https://cran.r-project.org/web/packages/Metrics/Metrics.pdf

27. Lee J-MM, Kim Y-WY, Welk GJ. Validity of consumer-based physical activity monitors. Med Sci Sports Exerc [Internet]. United States, United States; 2014 [cited 2017 Nov 20];46:1840–8. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=sph&AN=96789748&site=ehost-live

28. Team R. Core. R: A language and environment for statistical computing. 2018;0:201.

29. Wright SP, Hall Brown TS, Collier SR, Sandberg K. How consumer physical activity monitors could transform human physiology research. Am J Physiol Integr Comp Physiol [Internet]. S.P. Wright, Georgetown Univ., 232 Building D, 4000 Reservoir Rd., NW,

Washington, DC 20057, United States. E-mail: spw44@georgetown.edu, United States: American Physiological Society (E-mail: subscrip@the-aps.org); 2017 [cited 2017 Nov 17];312:R358–67. Available from: http://ajpregu.physiology.org/lookup/doi/10.1152/ajpregu.00349.2016

30. Ceesay SM, Prentice AM, Day KC, Murgatroyd PR, Goldberg GR, Scott W, et al. The use of heart rate monitoring in the estimation of energy expenditure : a validation study using indirect whole-body calorimetry. Brirish J Nutr [Internet]. England; 1989 [cited 2017 Oct 23];61:175–86. Available from: http://www.journals.cambridge.org/abstract_S0007114589000267

31. Leonard WR. Measuring human energy expenditure: What have we learned from the flex-heart rate method? Am J Hum Biol. 2003;15:479–89.

32. Gillinov S, ETIWY M, Wang R, Blackburn G, PHELAN D, Gillinov AM, et al. Variable Accuracy of Wearable Heart Rate Monitors during Aerobic Exercise. Med Sci Sport Exerc [Internet]. Lippincott Williams and Wilkins; 2017 [cited 2017 Nov 17];49:1697–703. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28709155

33. Rennie KL, Hennings SJ, Mitchell J, Wareham NJ. Estimating energy expenditure by heart-rate monitoring without individual calibration. Med Sci Sports Exerc [Internet]. 2001;33:939–45. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11404659&retmode=ref&cmd=prlinks%5Cnpapers2://publication/uuid/5CD2363C-779F-4417-B26D-4C2A80457601

34. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning The Elements of Statistical Learning. 2017; Available from: https://web.stanford.edu/~hastie/Papers/ESLII.pdf

35. Hopkins M, Finlayson G, Duarte C, Whybrow S, Ritz P, Horgan GW, et al. Modelling the associations between fat-free mass, resting metabolic rate and energy intake in the context

of total energy balance. Int J Obes [Internet]. Nature Publishing Group; 2016;40:312–8. Available from: http://dx.doi.org/10.1038/ijo.2015.155

36. Chowdhury EA, Western MJ, Nightingale TE, Peacock OJ, Thompson D. Assessment of laboratory and daily energy expenditure estimates from consumer multisensor physical activity monitors. PLoS One [Internet]. United States, United States: Public Library of Science (E-mail: plos@plos.org); 2017 [cited 2017 Nov 9];12:e0171720. Available from: http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0171720&type=printable

37. Lee J-M, Kim Y-W, Welk GJ. Validity and Utility of Consumer-Based Physical Activity. ACSM's Heal Fit J [Internet]. 2014;18:16–21. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=sph&AN=96789748&site=ehost-live

38. Wahl Y, Düking P, Droszez A, Wahl P, Mester J, Y. W, et al. Criterion-validity of commercially available physical activity tracker to estimate step count, covered distance and energy expenditure during sports conditions. Front Physiol [Internet]. Y. Wahl, Institute of Biomechanics and Orthopedics, German Sport University Cologne, Cologne, Germany. E-mail: y.wahl@dshs-koeln.de, Switzerland: Frontiers Media S.A. (E-mail: info@frontiersin.org); 2017 [cited 2017 Nov 13];8:725. Available from: http://journal.frontiersin.org/article/10.3389/fphys.2017.00725/full

39. Slinde F, Bertz F, Winkvist A, Ellegård L, Olausson H, Brekke HK. Energy expenditure by multisensor armband in overweight and obese lactating women validated by doubly labeled water. Obesity [Internet]. United States; 2013 [cited 2017 Dec 10];21:2231–5. Available from: http://doi.wiley.com/10.1002/oby.20363

40. Casiraghi F, Lertwattanarak R, Luzi L, Chavez AO, Davalli AM, Naegelin T, et al. Energy Expenditure Evaluation in Humans and Non-Human Primates by SenseWear Armband. Validation of Energy Expenditure Evaluation by SenseWear Armband by Direct

Comparison with Indirect Calorimetry. PLoS One [Internet]. F. Folli, Department of Medicine/Division of Diabetes, University of Texas Health Science Center at San Antonio, San Antonio, TX, United States. E-mail: folli@uthscsa.edu, United States: Public Library of Science (185 Berry Street, Suite 1300, San Francisco CA 94107, United States); 2013;8:e73651. Available from: http://www.plosone.org/article/fetchObject.action

41. Johannsen DL, Calabro MA, Stewart J, Franke W, Rood JC, Welk GJ. Accuracy of armband monitors for measuring daily energy expenditure in healthy adults. Med Sci Sports Exerc [Internet]. G. J. Welk, 257 Forker Bldg., Ames, IA 50011, United States. E-mail: gwelk@iastate.edu: Lippincott Williams and Wilkins (530 Walnut Street,P O Box 327, Philadelphia PA 19106-3621, United States); 2010;42:2134–40. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=sph&AN=55265135&site=ehost-live

42. Shook RP, Hand GA, O'Connor DP, Thomas DM, Hurley TG, Hébert JR, et al. Energy Intake Derived from an Energy Balance Equation, Validated Activity Monitors, and Dual X-Ray Absorptiometry Can Provide Acceptable Caloric Intake Data among Young Adults. J Nutr [Internet]. United States; 2018;148:490–6. Available from: https://academic.oup.com/jn/article/148/3/490/4930797

43. Sanghvi A, Redman LM, Martin CK, Ravussin E, Hall KD. Validation of an inexpensive and accurate mathematical method to measure long-term changes in free-living energy intake. Am J Clin Nutr. United States; 2015;102:353–8.

44. Thomas DDM, Schoeller DA, Redman LA, Martin CK, Levine JA, Heymsfield SB. A computational model to determine energy intake during weight loss. Am J … [Internet]. United States; 2010;92:1326–31. Available from: http://ajcn.nutrition.org/content/92/6/1326.short

45. Schoeller DA, Ravussin E, Schutz Y, Acheson KJ, Baertschi P, Jequier E, et al. Energy expenditure by doubly labeled water: validation in humans and proposed calculation. Am J

Physiol - Regul Integr Comp Physiol [Internet]. 1986;250:R823–30. Available from: http://ajpregu.physiology.org/content/250/5/R823.abstract

46. Byrne NM, Hills AP, Hunter GR, Weinsier RL, Schutz Y. Metabolic equivalent: one size does not fit all. J Appl Physiol [Internet]. 2005 [cited 2019 May 7];99:1112–9. Available from: http://www.physiology.org/doi/10.1152/japplphysiol.00023.2004

Tables

Table 1: Predictive variables included in each of the random forest models.

Fitbit Charge 2		
	Model	
	$FBRF_1$	**Device outputs:** Fitbit HR above sitting HR, Steps, **Subject characteristics:** Age, Gender, Height, weight
	$FBRF_2$	1 + FM (kg), FFM (kg)
Sensewear Armband Mini		
	$SWRF_1$	**Device outputs:** Average (Axis: X, Y, Z), Peaks (X, Y, Z), Mean absolute deviation (X, Y, Z), steps/min-1, Near body temperature average, skin temperature average, Galvanic skin response. **Subject characteristics:** age, gender, Height, weight,
	$SWRF_2$	**1 +** polar HR above sitting HR
	$SWRF_3$	2 + FM(kg), FFM (kg)
Actigraph GT3-x	$AGRF_1$	**Time domain, multi-axis (X, Y, Z) and first order differential (XYZ) features:**

		minimum, maximum, mean, standard deviation, correlation (XY, XZ, YZ), Median 0 crossings, percentiles (10, 25, 50, 75, 90th) **Frequency domain multi-axis (X, Y, Z) and first order differential (XYZ) features:** dominant frequency, dominant frequency magnitude **Subject characteristics:** Gender, age, height, weight,
	$AGRF_2$	**1 +** Polar Heart rate above sitting HR
	$AGRF_3$	2 + FM (kg), FFM (kg)

Abbreviations: Fitbit random forest (FBRF), Sensewear random forest (SWRF), Actigraph random forest (AGRF).

Table 2: Accuracy statistics computed using the LOOCV validation approach. Criterion METs Refers to METs calculated from indirect calorimetry, Predicted METs refers to model prediction. Minutes pooled refers to the number of minutes used for validation and participants refers to the number of participants included in each validation. RMSE, MAPE and correlation are presented with 95% confidence intervals. Equivalence refers to the results of the equivalence tests.

	Minutes	Participants	Criterion METs	Predicted METs	RMSE (METs)	MAPE (%)	Correlation (r)	Equivalence
SWA	2188	59	4.19 ± 2.61	3.72 ± 2.40	1.8	33.57	0.76	not equivalent
AGRF$_1$	2161	58	4.20 ± 2.61	4.18 ± 2.32	1.18	20.9	0.89	equivalent
AGRF$_2$	2125	57	4.22 ± 2.63	4.21 ± 2.37	1.03	18.31	0.92	equivalent
AGRF$_3$	2049	55	4.22 ± 2.62	4.21 ± 2.37	1.02	18.52	0.92	equivalent
FBRF$_1$	2077	57	4.19 ± 2.63	4.16 ± 2.30	1.36	28.74	0.86	equivalent
FBRF$_2$	2001	55	4.19 ± 2.63	4.14 ± 2.20	1.37	30.59	0.85	equivalent
SWRF$_1$	2188	59	4.19 ± 2.61	4.19 ± 2.24	1.24	23.61	0.88	equivalent
SWRF$_2$	2153	58	4.21 ± 2.62	4.22 ± 2.35	1	18.82	0.92	equivalent
SWRF$_3$	2077	56	4.21 ± 2.62	4.22 ± 2.34	1.02	19.38	0.92	equivalent

Abbreviations: Metabolic equivalents (METs), Root mean squared error (RMSE), Mean absolute percentage error (MAPE) , Actigraph random forest (AGRF), Fitbit random forest (FBRF), Sensewear random forest (SWRF).

Table 3: Accuracy statistics computed using the LOOCV validation approach, presented for each of the activities performed. Criterion METs Refers to METs calculated from indirect calorimetry, Predicted METs refers to model prediction. Minutes pooled refers to the number of minutes used for validation and participants refers to the number of participants included in each validation. RMSE, MAPE and correlation are presented with 95% confidence intervals. Equivalence refers to the results of the equivalence tests.

Activity	Model	Minutes pooled	Participants	Criterion METs	Predicted METs	RMSE (METs)	MAPE (%)	Correlation (r)	Equivalence
Sit									
	SWA	236	59	1.18 ± 0.21	1.15 ± 0.23	0.27	15.03	0.26	equivalent
	AGRF 1	232	58	1.18 ± 0.21	1.33 ± 0.41	0.44	22.43	0.21	not equivalent
	AGRF 2	227	57	1.18 ± 0.21	1.28 ± 0.26	0.31	18.77	0.25	not equivalent
	AGRF 3	219	55	1.18 ± 0.21	1.28 ± 0.26	0.31	19.38	0.25	not equivalent
	FBRF1	223	57	1.18 ±0.21	1.49 ± 0.22	0.43	31.95	0.02	not equivalent
	FBRF2	215	55	1.18 ±0.21	1.63 ± 0.22	0.54	42.92	-0.06	not equivalent
	SWRF 1	236	59	1.18 ± 0.21	1.35 ± 0.31	0.38	22.2	0.19	not equivalent
	SWRF 2	231	58	1.18 ± 0.21	1.27 ± 0.21	0.27	16.98	0.27	equivalent
	SWRF 3	223	56	1.18 ± 0.21	1.27 ±0.21	0.27	17.64	0.26	not equivalent
Stand									
	SWA	236	59	1.34 ± 0.24	1.34 ± 0.43	0.49	29.92	0.04	equivalent
	AGRF 1	232	58	1.34 ± 0.25	1.41 ± 0.43	0.44	19.95	0.27	equivalent
	AGRF 2	227	57	1.34 ± 0.24	1.4 ± 0.33	0.36	18.08	0.28	equivalent
	AGRF 3	219	55	1.34 ± 0.24	1.4 ± 0.34	0.36	18.38	0.29	equivalent
	FBRF1	225	57	1.34 ± 0.25	1.8 ± 0.63	0.84	47.87	0.04	not equivalent
	FBRF2	217	55	1.33 ± 0.25	1.93 ± 0.5	0.82	52.3	-0.05	not equivalent
	SWRF 1	236	59	1.34 ± 0.24	1.57 ± 0.65	0.71	31.83	0.07	not equivalent
	SWRF 2	231	58	1.34 ± 0.24	1.44 ± 0.4	0.46	21.82	0.11	not equivalent
	SWRF 3	223	56	1.34 ± 0.24	1.45 ± 0.42	0.47	22.74	0.11	not equivalent
Walk									
	SWA	236	59	3.95 ± 0.61	3.6 ± 0.45	0.91	17.98	-0.25	not equivalent
	AGRF 1	232	58	3.94 ± 0.61	4.54 ± 0.55	0.99	22.23	0.06	not equivalent
	AGRF 2	228	57	3.94 ± 0.62	4.26 ± 0.6	0.83	17.05	0.19	not equivalent
	AGRF 3	220	55	3.97± 0.61	4.28 ± 0.6	0.84	17.04	0.19	not equivalent
	FBRF1	219	57	3.93 ± 0.6	4.28 ± 0.39	0.79	17.84	0.02	not equivalent
	FBRF2	211	55	3.95 ± 0.6	4.28 ± 0.4	0.78	16.8	0.03	not equivalent
	SWRF 1	236	59	3.94 ± 0.61	4.55 ± 0.38	0.89	20.2	0.19	not equivalent
	SWRF 2	232	58	3.95 ± 0.61	4.25 ± 0.52	0.69	15.34	0.4	equivalent
	SWRF 3	224	56	3.97 ± 0.61	4.29 ± 0.52	0.69	15.18	0.42	equivalent
Walk incline									

	SWA	236	59	5.24 ± 0.72	4.15 ± 0.85	1.65	26.14	-0.26	not equivalent
	AGRF 1	232	58	5.23 ± 0.72	4.68 ± 0.58	1.03	14.14	0.11	not equivalent
	AGRF 2	228	57	5.23 ± 0.73	4.82 ±0.65	0.96	13.27	0.21	equivalent
	AGRF 3	220	55	5.26 ± 0.72	4.85 ± 0.66	0.96	13.01	0.21	equivalent
	FBRF1	216	57	5.19 ± 0.72	4.36 ± 0.46	1.16	17.01	0.1	not equivalent
	FBRF2	208	55	5.22 ± 0.72	4.33 ± 0.49	1.2	18.87	0.14	not equivalent
	SWRF 1	236	59	5.24 ± 0.73	4.92 ± 0.97	1.28	18.36	-0.05	equivalent
	SWRF 2	232	58	5.24 ± 0.73	4.97 ±0.65	0.94	13.86	0.14	equivalent
	SWRF 3	224	56	5.27 ± 0.72	5 ± 0.65	0.93	13.48	0.14	equivalent
Jog									
	SWA	195	49	8.57 ± 1.21	8.12 ± 1.52	2.01	19.3	-0.01	equivalent
	AGRF 1	195	49	8.57 ± 1.21	8.73 ± 1.12	1.54	15.87	0.14	equivalent
	AGRF 2	195	49	8.57 ± 1.21	8.63 ± 1.15	1.38	13.81	0.32	equivalent
	AGRF 3	187	47	8.54 ± 1.23	8.59 ± 1.14	1.33	13.56	0.36	equivalent
	FBRF1	191	48	8.55 ± 1.22	8.68 ± 1.19	1.5	15.3	0.22	equivalent
	FBRF2	183	46	8.52 ± 1.24	8.46 ±1.18	1.54	15.57	0.18	equivalent
	SWRF 1	195	49	8.57 ± 1.21	8.48 ± 1.15	1.34	12.33	0.36	equivalent
	SWRF 2	195	49	8.57 ± 1.21	8.59 ± 1.08	1.15	11.26	0.5	equivalent
	SWRF 3	187	47	8.54 ± 1.23	8.5 ±1.1	1.17	11.45	0.5	equivalent
Jog incline									
	SWA	120	30	10.07 ± 1.32	8.04 ± 1.75	3.03	24.69	-0.06	not equivalent
	AGRF 1	120	30	10.07 ± 1.32	8.59 ± 1.43	2.42	19.33	0.03	not equivalent
	AGRF 2	120	30	10.07 ± 1.32	8.86 ±1.45	2.19	18.03	0.13	not equivalent
	AGRF 3	116	29	10.06 ± 1.34	8.98 ±1.37	2.04	17.12	0.17	not equivalent
	FBRF1	120	30	10.07 ± 1.32	8.86 ±1.54	2.1	15.9	0.28	not equivalent
	FBRF2	116	29	10.06 ± 1.34	8.7 ±1.5	2.21	17.4	0.24	not equivalent
	SWRF 1	120	30	10.07 ± 1.32	8.36 ± 1.48	2.45	19.15	0.21	not equivalent
	SWRF 2	120	30	10.07 ± 1.32	8.88 ±1.24	1.83	15.2	0.41	not equivalent
	SWRF 3	116	29	10.06 ± 1.34	8.9 ± 1.26	1.81	15.03	0.42	not equivalent
Cycle low									
	SWA	233	59	4.15 ± 1.08	2.53 ± 0.96	1.93	40.49	0.46	not equivalent
	AGRF 1	229	58	4.13 ± 1.07	4.33 ±0.74	1.25	25.11	0.11	equivalent
	AGRF 2	225	57	4.14 ± 1.07	4.33 ±0.76	1.03	20.8	0.42	equivalent
	AGRF 3	217	55	4.14 ± 1.08	4.32 ±0.75	1.06	21.6	0.39	equivalent
	FBRF1	220	56	4.11 ±1.09	3.77 ±1.08	1.54	29.31	0.03	not equivalent
	FBRF2	212	54	4.1 ± 1.1	3.63 ± 0.92	1.45	26.67	0.09	not equivalent
	SWRF 1	233	59	4.15 ± 1.08	4.25 ±0.75	1.11	22.16	0.3	equivalent

	SWRF 2	229	58	4.17 ± 1.08	4.23 ± 0.86	0.93	17.81	0.56	equivalent
	SWRF 3	221	56	4.17 ± 1.09	4.2 ± 0.86	0.97	18.39	0.53	equivalent
Cycle mid									
	SWA	225	58	5.21 ± 1.43	3.31 ± 1.75	2.55	41.99	0.44	not equivalent
	AGRF 1	225	58	5.21 ± 1.43	4.53 ± 0.65	1.61	20.11	0.18	not equivalent
	AGRF 2	221	57	5.22 ± 1.44	4.78 ± 0.81	1.32	16.67	0.5	not equivalent
	AGRF 3	213	55	5.25 ±1.46	4.82 ± 0.83	1.34	17.08	0.5	not equivalent
	FBRF1	207	55	5.19 ± 1.44	3.71 ± 1.13	2.27	30.96	0.13	not equivalent
	FBRF2	199	53	5.22 ± 1.47	3.61 ± 0.94	2.27	30.75	0.16	not equivalent
	SWRF 1	225	58	5.21 ± 1.43	4.16 ± 0.84	1.79	22.21	0.26	not equivalent
	SWRF 2	221	57	5.22 ± 1.44	4.59 ± 1.03	1.47	18.89	0.47	not equivalent
	SWRF 3	213	55	5.25 ± 1.46	4.58 ± 1.01	1.49	19.03	0.46	not equivalent
Folding									
	SWA	236	59	2.75 ± 0.6	4.29 ± 1.77	2.37	72.37	0.12	not equivalent
	AGRF 1	232	58	2.75 ± 0.6	2.91 ± 0.35	0.61	18.98	0.32	equivalent
	AGRF 2	228	57	2.75 ± 0.61	2.95 ± 0.38	0.59	18.8	0.44	equivalent
	AGRF 3	220	55	2.75 ± 0.61	2.96 ± 0.39	0.63	19.75	0.36	equivalent
	FBRF1	228	57	2.74 ± 0.61	3.5 ± 0.71	1.19	37.21	0.05	not equivalent
	FBRF2	220	55	2.75 ± 0.61	3.58 ± 0.71	1.26	39.8	-0.01	not equivalent
	SWRF 1	236	59	2.75 ± 0.6	3.43 ± 0.35	1.04	33.41	0.08	not equivalent
	SWRF 2	232	58	2.74 ± 0.6	3.37 ± 0.69	0.98	30.06	0.32	not equivalent
	SWRF 3	224	56	2.75 ± 0.61	3.42 ± 0.7	1.03	32.08	0.29	not equivalent
Sweeping									
	SWA	235	59	3.12 ± 0.71	3.52 ± 1.47	1.59	41.49	0.13	not equivalent
	AGRF 1	232	58	3.13 ± 0.71	3.64 ± 0.68	1.03	29.33	0.17	not equivalent
	AGRF 2	226	57	3.12 ± 0.71	3.57 ± 0.68	0.95	27.09	0.28	not equivalent
	AGRF 3	218	55	3.13 ± 0.71	3.56 ± 0.63	0.93	27	0.24	not equivalent
	FBRF1	228	57	3.13 ± 0.72	3.95 ± 0.67	1.2	35.29	0.21	not equivalent
	FBRF2	220	55	3.13 ± 0.71	3.96 ± 0.69	1.22	35.33	0.2	not equivalent
	SWRF 1	235	59	3.13 ± 0.72	3.65 ± 0.83	1.14	30.04	0.15	not equivalent
	SWRF 2	230	58	3.11 ± 0.71	3.51 ± 0.92	0.97	24.11	0.43	not equivalent
	SWRF 3	222	56	3.12 ± 0.71	3.57 ± 0.95	1.02	25.41	0.42	not equivalent

Abbreviations: Metabolic equivalents (METs), Root mean squared error (RMSE), Mean absolute percentage error (MAPE), Fitbit random forest (FBRF), Sensewear random forest (SWRF), Actigraph random forest (AGRF).

Legends for figures

Figure 1. A scatter plot of the Measured METs (Vyntus CPX) and predicted METs. Diagonal lines represent lines of identity. Data are shown for the Actigraph random forest model ($AGRF_2$, top left), Fitbit random forest model ($FBRF_1$, top right), Sensewear armband (SWA, bottom left) and Sensewear armband random forest ($SWRF_2$, bottom right).

Figure 2. Variable importance plots detailing the increase in mean squared error associated with permutation (random shuffling) of a single variable. Data are shown in order of importance for the first 20 variables. Data are shown for models: $FBRF_2$ (blue dots, left), $SWRF_3$ (red dots, middle), and $AGRF_3$ (green dots, right).
Abbreviations: HR = Heart rate, FM (kg)= Fat mass (kg), FFM (KG) = Fat free mass (kg), MAD = Mean absolute deviation, FOD = First order differential.