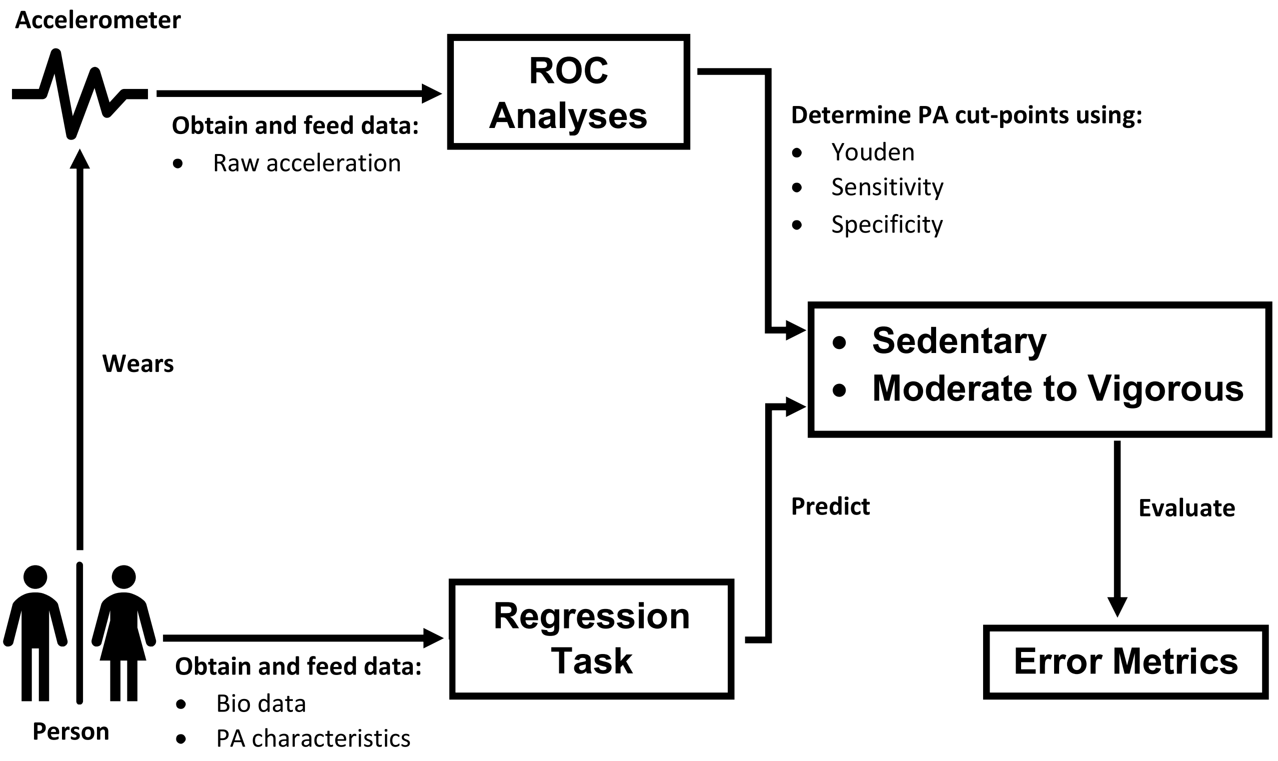
Graphical Abstract

Personalised Accelerometer Cut-point Prediction for Older Adults' Movement Behaviours using a Machine Learning approach

Nonso Nnamoko,Luis Adrián Cabrera-Diego,Daniel Campbell,George Sanders,Stuart J. Fairclough,Ioannis Korkontzelos



Highlights

Personalised Accelerometer Cut-point Prediction for Older Adults' Movement Behaviours using a Machine Learning approach

Nonso Nnamoko,Luis Adrián Cabrera-Diego,Daniel Campbell,George Sanders,Stuart J. Fairclough,Ioannis Korkontzelos

- A model is developed to predict physical activity cut-points on accelerometer based on individual characteristics
- Post data collection analytical process helps towards a standardised method for characterising physical activity
- Multiple features calculated from raw accelerometer data was used to enrich the feature set for training machine learning
- Personalisation was achieved by combining raw accelerometer data with person-specific data e.g., blood pressure

Personalised Accelerometer Cut-point Prediction for Older Adults' Movement Behaviours using a Machine Learning approach[⋆,⋆⋆]

Nonso Nnamoko[a,*], Luis Adrián Cabrera-Diego[a,c], Daniel Campbell[a,e], George Sanders[d], Stuart J. Fairclough[b] and Ioannis Korkontzelos[a,**]

[a] Department of Computer Science, Edge Hill University, Ormskirk, L39 4QP, United Kingdom
[b] Department of Sports and Physical Activity, Edge Hill University, Ormskirk, L39 4QP, United Kingdom
[c] Faculté des Sciences et Technologies, La Rochelle Université, La Rochelle, 17042, France
[d] Carnegie School Of Sport, Leeds Beckett University, Leeds, LS1 3HE, United Kingdom
[e] School of Physical Sciences and Computing, University of Central Lancashire, Preston, PR1 2HE, United Kingdom

ARTICLE INFO

Keywords:
Physical activity
Energy expenditure
Machine Learning
Accelerometry

Abstract

Background and Objectives: Body-worn accelerometers are the most popular method for objectively assessing physical activity in older adults. Many studies have developed *generic* accelerometer cut-points for defining activity intensity in metabolic equivalents for older adults. However, methodological diversity in current studies has led to a great deal of variation in the resulting cut-points, even when using data from the same accelerometer. In addition, the *generic* cut-point approach assumes that 'one size fits all' which is rarely the case in real life. This study proposes a machine learning method for personalising activity intensity cut-points for older adults.

Methods: Firstly, raw accelerometry data was collected from 33 older adults who performed set activities whilst wearing two accelerometer devices: GENEActive (wrist worn) and ActiGraph (hip worn). ROC analysis was applied to generate *personalised* cut-point for each data sample based on a device. Four cut-points have been considered: Sensitivity optimised Sedentary Behaviour; Specificity optimised Moderate to Vigorous Physical Activity; Youden optimised Sedentary Behaviour; and Youden optimised Moderate to Vigorous Physical Activity. Then, an *additive* regression algorithm trained on biodata features, that concern the individual characteristics of participants, was used to predict the cut-points. As the model output is a numeric cut-point value (and not discrete), evaluation was based on two error metrics, Mean Absolute Error and Root Mean Square Error. Standard Error of estimation was also calculated to measure the accuracy of prediction (goodness of fit) and this was used for performance comparison between our approach and the state-of-the-art. Hold-out and 10-Fold cross validation methods were used for performance validation and comparison.

Results: The results show that our *personalised* approach performed consistently better than the state-of-the-art with 10-Fold cross validation on all four cut-points considered for both devices. For the ActiGraph device, the Standard Error of estimation from our approach was lower by 0.33 (Youden optimised Sedentary Behaviour), 9.50 (Sensitivity optimised Sedentary Behaviour), 0.64 (Youden optimised Moderate to Vigorous Physical Activity) and 22.11 (Specificity optimised Moderate to Vigorous Physical Activity). Likewise, the Standard Error of estimation from our approach was lower for the GENEActiv device by 2.29 (Youden optimised Sedentary Behaviour), 41.65 (Sensitivity optimised Sedentary Behaviour), 4.31 (Youden optimised Moderate to Vigorous Physical Activity) and 347.15 (Specificity optimised Moderate to Vigorous Physical Activity).

Conclusions: *personalised* cut-point can be predicted without prior knowledge of accelerometry data. The results are very promising especially when we consider that our method predicts cut-points without prior knowledge of accelerometry data, unlike the state-of-the-art. More data is required to expand the scope of the experiments presented in this paper.

1. Introduction

Objectively representing daily physical activity (PA) is crucial, particularly in studies that involve older adults, where increasing PA and/or reducing sedentary behaviour (SB) is often the intended outcome [5]. Among the numerous devices available for measuring PA, body-worn accelerometers are the least obtrusive, thus are increasingly utilised for this purpose [40, 62, 64, 45]. However, there are arguments against the validity of the results in calibration studies involving older adults. This is in part because the underlying standards used to determine metabolic costs are not applicable to older adults [4]. Furthermore, the methods used to translate accelerometer outputs into activity intensity thresholds are too diverse [42, 58].

[⋆] This document is the results of an extended research using data from Sanders et al.

[⋆⋆] The tools and methods used in this research are adapted from the TYPHON Project, funded by the EU Horizon 2020 Programme under grant No. 780251.

[*] Corresponding author
[**] Principal corresponding author
nnamokon@edgehill.ac.uk (N. Nnamoko); luis.cabrera_diego@univ-lr.fr (L.A. Cabrera-Diego); campbeld@edgehill.ac.uk (D. Campbell); G.Sanders@leedsbeckett.ac.uk (G. Sanders); Stuart.Fairclough@edgehill.ac.uk (S.J. Fairclough); Yannis.Korkontzelos@edgehill.ac.uk (I. Korkontzelos)
ORCID(s): 0000-0002-5064-2621 (N. Nnamoko); 0000-0002-9881-9799 (L.A. Cabrera-Diego); 0000-0003-4244-9458 (D. Campbell); 0000-0001-8358-1979 (S.J. Fairclough); 0000-0001-8052-2471 (I. Korkontzelos)

Generally, PA recommendations for health benefits are intensity specific; typically categorised into light, moderate and vigorous intensity based on metabolic equivalents (MET) [27]. One MET equates to the standard resting metabolic rate (RMR), i.e., the oxygen (O_2) consumption required at rest or sitting quietly, and in healthy adults is assumed to be 3.5 mL $\times$ kg^{-1} $\times$ min^{-1}. This index is used to express O_2 uptake or activity intensity in multiples of the value of 1 MET and is useful for estimating and prescribing exercise of different intensities. For example, activities may range from sleeping (0.9 MET) to running at 10.9 mph (18 METs) [52]; and 3 METs represent the commonly-accepted cut-off value between light and moderate intensity PA [1]. Currently, this index is commonly used for categorising PA intensity in observational studies for older adults [3, 6, 38, 53]. However, the actual energy cost varies between individuals due to differences in body mass, age, health status etc. and it is well established that RMR decreases with age [33, 9, 20, 30, 39]. Thus, for individuals of different size and age, the energy expenditure estimates are influenced by the consistency of the assumed RMR value of 3.5 mL $\times$ kg^{-1} $\times$ min^{-1}. In other words, computing MET-based PA intensity values using the RMR index has implications for older adults. With under-estimated energy expenditure, older adults would be exercising at higher relative intensities than assumed and their time spent in PA above activity intensity thresholds would be under-estimated [21]. Therefore, it is not surprising that a growing number of research studies have found this index to be inaccurate across individuals with heterogeneous physical, demographic and health status characteristics [9, 60].

Another growing concern is that there is no standardised method to translate accelerometer output into an estimate of physical activity for older adults [41]. This is in part due to methodological diversity in the energy expenditure equations used in existing calibration studies to translate accelerometer output into measures of MET expenditure that reflect thresholds for specified levels of PA [58, 42]. The methodological diversity of these studies has produced a wide variety of predictive equations and cut-points for PA, even when assessing the same accelerometer. This diversity reduces the ability to interpret results obtained from same accelerometers, among different research studies or even among different accelerometer types. Consequently, research in the area is moving towards post-data collection analytical methods, such as supervised/unsupervised machine learning for for free living PA, rather than lab caliberation protocols. Such methods can be replicated easily, thereby providing greater methodological transparency and improve comparability between different studies and accelerometer models.

On that note, the research presented in this paper is motivated by a recent study by Sanders et al. [51] that used a post-data collection analytical process to estimate *generic* cut-points for SB and moderate to vigorous physical activity (MVPA) in older adults. Specifically, the study used the accelerometer output from GENEActiv[1] (GA) and ActiGraph[2] (AG) obtained from a heterogeneous sample of 34 older adults (mean age = 69.6, SD = 8.0) to determine raw acceleration cut-points for SB and MVPA. GA is worn on the non dominant wrist while AG is attached to the left hip area. Two approaches based on receiver operative characteristic (ROC) curve analysis [29] were adopted to achieve this. The results were promising but the 'one size fits all' approach that is known to produce inconsistent result across individuals of different body mass and age [9, 60] was maintained.

In this paper, we go a step further by proposing a model capable of personalising raw acceleration cut-points for SB and MVPA for older adults, according to their individual characteristics. Specifically, we constructed an *additive* regression model [19] that describes the relationship between aspects of an individual, i.e., *input features* such as age, gender, weight etc., and the value of interest, i.e., *output features* such as cut-points for SB or MVPA). The model can generate estimated outputs when given a new set of *input features*. All experiments were conducted with the same data used in Sanders et al. [51], so that the results are fully comparable.

The model predicts values within a range rather than discrete class labels, e.g., 'MVPA' or 'not MVPA'. Thus, the accuracy of such model is typically evaluated via the error in the predicted values [61]. We used the mean absolute error (MAE) and the root mean squared error (RMSE) as metrics. Standard Error of estimation was also calculated to measure the accuracy of prediction i.e., 'goodness of fit' of the regression models and this was used for performance comparison between our approach and the state-of-the-art which we used as the *Baseline*. As an evaluation *Baseline*, we used the results published in Sanders et al. [51], which is the only study that uses a post data-collection process to determine raw acceleration cut-points for older adults. The results suggest that *generic* cut-points are unreliable. The proposed *personalised* approach is a superior alternative to the state-of-the-art as proven by the results which shows higher performance consistently across the cut-points considered in this study. There is also room for improvement especially given a larger training data.

This study makes the following contributions:

i Post data collection analysis using machine learning to predict *personalised* PA acceleration cut-points for older adults:
To the best of our knowledge, this is the first study to personalise activity classification thresholds among this age group using a standardised approach.

ii A priori rather than a posteriori PA acceleration cut-point determination:
We predict cut-points for SB and MVPA based on the general characteristics of a person such as age, gender, weight etc. This contradicts the state-of-the-art, where

[1]www.activinsights.com
[2]www.actigraphcorp.com

accelerometry data is known and used to calibrate cut-points.

iii *personalised* PA acceleration cut-points is feasible and superior:
The results presented in this paper indicates that *personalised* PA acceleration cut-points is a feasible and superior alternative to *generic* ones. The personalisation approach presented in this paper prove absolute superiority over the state-of-the-art. We believe that a larger training data would lead to further improvement and thus, a result that can be generalised.

The remainder of this paper is organised as follows: A concise overview of related work is provided in Section 2, followed by a detailed explanation of the experimental data in Section 3.1. In Section 3.2, we discuss our approach. This is followed by the experimental setup in Section 3.3, and experimental results in Section 4. We put the results into context in Section 5 before Section 6, where we present a summary of conclusions drawn from the entire experiments as well as recommendations based on the findings.

2. State-of-the-art

The number of older adults continues to grow at an unprecedented rate globally, with individuals who are 60 years or older accounting for 8.5% (617 million) of the populace in 2016; and projected to rise to 17% (1.6 billion) by 2050 [22]. Such increase in ageing population presents several public health challenges, thus positive lifestyle is often encouraged among older adults to maintain good health, functionality and independent living. Body-worn accelerometers offer an objective means to assess free-living physical activity by measuring movement. They are capable of sensing and recording unfiltered movement activity, which can then be used to determine time spent in SB and/or MVPA. The most common means of doing so is to translate accelerometer output into measures of MET expenditure that map to thresholds for specified levels of physical activity [58]. Although numerous studies have attempted to calibrate and validate accelerometers, there is no standardised method to translate accelerometer output into an estimate of physical activity for older people [41]. The majority of calibration studies for adults have typically developed prediction equations that use oxygen expenditure as a criterion measure to translate activity counts into measured activity intensity levels [42]. However, the wide range of methods used in these studies has lead to great variation in the developed energy expenditure equations and the resultant activity intensity thresholds or cut-points calculated from them, even when using the same monitor [58].

Several studies [7, 18, 36, 8, 42, 56, 63, 24] have been conducted to examine the validity of ActiGraph, a uniaxial accelerometer popularly used in PA research, for measuring PA [55, 48, 47, 26, 13]. We summarised their results in Table 1 to show the various sample size; age and gender composition of the study population; the energy expenditure function used in each study; and the resultant PA acceleration cut-points deduced for defining moderate and vigorous activity. We observe that all the studies arrived at a radically different cut-points for moderate intensity activity (range:191–2743) and vigorous intensity activity (range: 4945–7526); even though they all used the same Actigraph accelerometer data.

A common theme among the studies in Table 1 is that gross energy expenditure predictive equations were used to translate accelerometer output into measures of MET expenditure so that cut-points can be determined. However, recent accelerometer-based PA research has moved toward post-data collection analytical methods involving raw acceleration rather than counts [51]. Since accelerometers are capable of recording raw unfiltered movement activity, researchers [59, 51, 50] believe that the data lends itself to further development of innovative metrics. On that note, Fairclough et al. [17] argued that accelerometer output can be post-processed by reducing the data to dimensionless activity "counts" per user-specified period of time or epoch. Then, the data can be processed with standardised methods, e.g., machine learning, to generate activity intensity threshold values. Hildebrand et al. [26] added that post-data collection analysis with standardised methods is likely to provide greater methodological transparency and improve comparability of results.

To the best of our knowledge, Sanders et al. [51] is the only study to use a standardised method to process raw acceleration, with a view to generate distinguishing cut-points for PA intensity in older adults. Specifically, the authors used receiver operative characteristic (ROC) curve analysis [29] to determine cut-points for SB and MVPA. The study was conducted with raw acceleration data from GA and AG devices obtained from a heterogeneous sample of 34 older adults (59 - 86 years old, 44 - 115kg, 10 male and 24 female); who engaged in a 2-visit laboratory PA protocol, that involved a mixture of ambulatory and lifestyle activities. Following recent studies that analysed body-worn accelerometer data [49, 43, 26, 17], Sanders et al. [51] used the Euclidean Norm Minus One (ENMO) [23] to quantify acceleration values from the devices, i.e., GA and AG, in relation to gravity (1 mg = 0.00981 ms^{-2}). The raw data was further reduced by averaging the ENMO values over 1-second epochs. Thus, the resulting ENMO values are expressed in milli (10^{-3}) gravity-based acceleration units (mg), where 1 g = 9.81 m/s^2.

The experiments took a *calibration* vs. *validation* approach in which a randomly counter-balanced sample of 17 participants (12 female, five male) from visit 1, and 17 participants (12 female, five male) from visit 2 was used for calibration and the rest for validation. Basically, ENMO values from both devices, i.e., GA and AG, were first labelled as either SB or MVPA. The activPAL[3] accelerometer worn by participants on the left anterior thigh was used to categorise ENMO values into SB or not SB. Likewise, 3 MET VO_2 val-

[3]http://www.palt.com/

Table 1
Summary of experimental studies conducted for Actigraph 7164. Age distribution is a range or a mean and standard deviation. M and F stand for male and female, respectively.

Study	Number of Participants	Characteristics Sex	Age	Energy Expenditure Function $[counts/min]$	PA cut-point moderate	vigorous
Brage et al. [7]	12	M	[23, 30]	$2.886 + 7.429 \times 10^{-4} \times \text{counts/min} - 0.02 \times VO_2$	1810	5850
Freedson et al. [18]	50	M/F	24 ± 4	$1.439008 + 7.95 \times 10^{-4} \times \text{counts/min}$	1316	5354
Leenders et al. [36]	28	M/F	24 ± 4	$2.240 + 6 \times 10^{-4} \times \text{counts/min}$	1952	5725
Brooks et al. [8]	72	M/F	[35, 45]	$2.240 + 6 \times 10^{-4} \times \text{counts/min}$	1267	6252
Heil et al. [24]	58	M/F	28 ± 6	$1.551 + 6.19 \times 10^{-4} \times \text{counts/min}$	2341	7187
Yngve et al. [63]	28	M/F	23 ± 3	$1.136 + 8.249 \times 10^{-4} \times \text{counts/min}$	2260	5896
Yngve et al. [63]	28	M/F	23 ± 3	$0.751 + 8.198 \times 10^{-4} \times \text{counts/min}$	2743	6403
Hendelman et al. [25]	25	M/F	[30, 50]	$2.922 + 4.09 \times 10^{-4} \times \text{counts/min}$	191	7526
Swartz et al. [55]	70	M/F	[19, 74]	$2.608 + 6.863 \times 10^{-4} \times \text{counts/min}$	574	4945
Troiano et al. [56]	4867	M/F	[6, 70+]	Weighted average of cut-points in [7, 18, 63]	2020	5999

ues were used as the criterion reference standard for MVPA. SB and MVPA were each coded as either 0 (behaviour occurring) or 1 (behaviour not occurring). It is important to note that the observed mean RMR of the study participants was 2.89 mL $\times$ kg^{-1} $\times$ min^{-1} and this was used to define 1 MET.

Calibration was based on ROC curve analysis [29] to determine SB and MVPA cut-points for each device. Specifically, two different pairs of cut-points were generated by analysing combinations of sensitivity (Se) and specificity (Sp) on the ROC curves. Firstly, ENMO values that indicates a compromise between Se and Sp (Youden index) [46] is calculated for both SB and MVPA values and used as one set of cut-point (SB_{Youden} and $MVPA_{Youden}$). Youden is a suitable metric used in cases where Se and Sp are equally important and is given by eq 1.

$$\text{Youden} = \text{Max}_c \left(\text{Se}_c + \text{Sp}_c\right) \tag{1}$$

where c is the optimal compromise point.

Secondly, a set of cut-points were determined (i) by emphasising Se over Sp for SB (SB_{Se}) to minimise the likelihood of classifying SB as PA, and (ii) by emphasising Sp over Se for MVPA ($MVPA_{Sp}$), to reduce the likelihood of misclassifying light PA as MVPA. Consequently, a total of 4 cut-points were computed for each device, i.e., SB_{Youden}, $MVPA_{Youden}$, SB_{Se} and $MVPA_{Sp}$.

Using the established cut-points, a validation analysis was performed with data from the 17 participants (12 female, five male) whose visit 1 data was not used for calibration analysis, and the 17 participants (12 female, five male) whose visit 2 data was not used for calibration (N = 34). Specifically, the ENMO values were categorised into SB/not SB and MVPA/not MVPA. Then, two-by-two (2 × 2) contingency tables were used to compare them with the calibrated cut-points. The calibrated cut-points for both devices are shown in Table 2.

The results are promising but still maintain a 'one size fits all' approach, which has been shown to produce inconsistent result across individuals of different body mass and age [9, 60]. Consequently, the aim of our study is to personalise the raw acceleration cut-points for SB and MVPA based on the individual characteristics of each participant.

Table 2
Calibrated cut-points for GA and AG expressed in mg

Device	SB_{Youden}	$MVPA_{Youden}$	SB_{Se}	$MVPA_{Sp}$
GA	≤ 20	≥ 32	≤ 57	≥ 104
AG	≤ 6	≥ 19	≤ 15	≥ 69

3. Method and Materials

This section presents the experimental method we used to achieve the research aims. This includes the experimental setup implemented as well as a detailed description of the experimental data and its characteristics.

3.1. Data

We obtained the exact data used in Sanders et al. [51] to determine cut-points for SB and MVPA in older adults. The dataset contains measurements collected from 34 older adults, who took part in a laboratory-based protocol consisting of 16 activities (see Table 1 in [51]). The authors used GA and AG activity monitors worn on the non dominant wrist and left hip respectively, to measure raw triaxial accelerations at 60 Hz during the protocol.

i $\geq$ 59 years of age

ii physically cleared for exercise using the modified Physical Activity Readiness Questionnaire [10, 11]

iii have the ability to walk briskly on a treadmill without assistance

iv not taking any medications that would influence energy expenditure or ability to perform ambulatory activity

We experimented with 33 out of the 34 samples, due to large amount of missing values in the sample of a female participant. A total of 21 features were adopted for each sample, and the data characteristics are shown in Table 3.

Some features correspond to quantities that were measured directly during the study, whereas others were calculated later from the directly measurable ones. For example, the top part of Table 3 includes simple biodata associated to

Table 3
Experimental data features and characteristics (N=33)

Variable	Min	Max	Mean ± SD
Gender	0.00	1.00	0.70 ± 0.47
Age (years)	59.00	86.00	69.27 ± 7.93
Weight (kg)	44.00	115.00	71.27 ± 17.62
Height (m)	1.45	1.82	1.64 ± 0.10
BMI (kg/m^2)	20.50	41.00	26.10 ± 4.66
Systolic BP (mm/Hg)	109.00	195.00	145.72 ± 21.49
Diastolic BP (mm/Hg)	62.00	109.00	85.15 ± 11.19
6MWT (km/h)	1.70	6.80	4.37 ± 1.39
Walk at 65% speed (km/h)	1.10	4.40	2.84 ± 0.89
Walk at 75% speed (km/h)	1.30	5.10	3.28 ± 1.04
Walk at 85% speed (km/h)	1.40	5.80	3.72 ± 1.18
RMR ($mL \cdot kg^{-1} \cdot min^{-1}$)	2.69	6.59	4.09 ± 0.79
Min. VO_2 (L/min)	0.05	0.23	0.14 ± 0.05
Max. VO_2 (L/min)	0.86	2.14	1.47 ± 0.32
Avg. VO_2 (L/min)	0.39	0.78	0.55 ± 0.10
Min. EE ($mL \cdot kg^{-1} \cdot min^{-1}$)	0.62	3.47	1.94 ± 0.65
Max. EE ($mL \cdot kg^{-1} \cdot min^{-1}$)	13.49	32.63	21.21 ± 4.11
Avg. EE ($mL \cdot kg^{-1} \cdot min^{-1}$)	5.57	10.46	7.86 ± 1.13
Min. METs	0.21	1.20	0.67 ± 0.22
Max. METs	4.65	11.25	7.31 ± 1.42
Avg. METs	1.92	3.61	2.71 ± 0.39

the study participants, e.g., weight, height, body mass index (BMI), systolic and diastolic blood pressure (BP). The middle part of Table 3 includes a 6 minute walk test (6MWT) [2] on a threadmill to measure maximum walk speed; as well as walking on a threadmill at 3 maximal percentage speeds (65, 75 and 85) individually calibrated from the 6MWT. The top and middle part of the table was directly adopted from Sanders et al. [51]. The bottom part of Table 3 refers to data derived from oxygen consumption (VO_2) during the laboratory-based PA protocol. VO_2 was directly measured breath-by-breath in 1 second intervals and its value was used to calculate energy expenditure (EE), which was then used to classify activity intensity in METs. We enriched the data by performing further calculations to determine the minimum, maximum and average values per participant.

3.2. Experimental Method

Sanders et al. [51] calculated *generic* ROC-induced cut-points through group calibration involving data from half of the study sample. The target was to determine cut-points SB_{Youden}, $MVPA_{Youden}$, SB_{Se}, $MVPA_{Sp}$ for each device, that generalise across all study participants and indeed older adults ($\geq$ 60) in general, regardless of their individual characteristics such as age, weight and height. In our approach, we started by performing a preliminary experiment using ROC analysis to generate SB_{Youden}, $MVPA_{Youden}$, SB_{Se} and $MVPA_{Sp}$ cut-points per individual based on ENMO values from each device.

To estimate *personalised* cut-points for classifying PA as SB or MVPA, we developed an *additive* regression model [19] that learns the relationship between the *input*

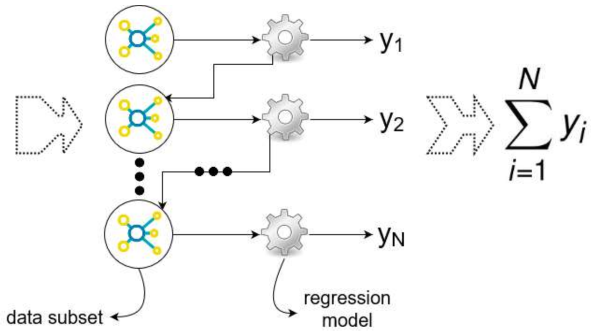


Figure 1: High level diagram of the experimental method

features in Table 3 and the *output features*, i.e., SB_{Youden}, $MVPA_{Youden}$, SB_{Se} and $MVPA_{Sp}$. The term 'additive' means the construction of a regression model in the form of an ensemble of *underlying* machine learners. This is done by sequentially fitting the *underlying* learner to the residuals left from the previous iteration. Specifically, at each iteration a subset of the training data is drawn at random (without replacement). The *underlying* learner is then trained on this randomly selected data and updates the prediction model for the current iteration. The final predicted values are computed by adding the predictions at each iteration. Figure 1 shows a high level diagram of the method.

We used *Decision Stump* [28] as the *underlying* machine learner for the *additive* regression model presented in this paper and we adopted ten (10) as the number of iterations. A *Decision Stump* is a one-level *Decision Tree* that makes prediction based on the value of a single input variable.

Unlike classification tasks that predict discrete classes, e.g., 'yes' or 'no', the output of the *additive* regression model is a measurable quantity, i.e., a numeric cut-point value. The performance of such a model is typically evaluated via the error in the predicted values [61]. Thus, we experimented with the mean absolute error (MAE) and the root mean squared error (RMSE) as metrics. These are well known error metrics commonly used to compare the performance of competing models. Since the model produces a numeric output given a set of *input variables*, MAE and RMSE allows to check the estimated output against the actual value that we tried to predict.

The MAE computes the average absolute difference between each actual output value, y_j, and the corresponding predicted value, $\hat{y}_j$:

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^{n} \left| y_j - \hat{y}_j \right| \tag{2}$$

where n is the total number of data points. To obtain the RMSE, we first calculated the mean square error (MSE) which measures the average squared difference between each actual output value, y_j, and the corresponding predicted value, $\hat{y}_j$:

$$\text{MSE (model)} = \frac{1}{n} \sum_{j=1}^{n} \left(y_j - \hat{y}_j \right)^2 \tag{3}$$

where n is the total number of data points. The RMSE is the square root of MSE.

$$\text{RMSE} = \sqrt{\frac{\sum_{j=1}^{n} \left(y_j - \hat{y}_j\right)^2}{n}} \tag{4}$$

For every data point, MAE and RMSE condense the differences between each actual and predicted value into a single value, and represent the predictive ability of the model. Each predicted value is expected to be off from the actual value by no more than the MAE or RMSE, whichever is used for evaluation. For example, if y_j is a MVPA actual *output value*, the predicted value $\hat{y}_j$ is considered correct if y_j value is greater or equal to y_j−MAE or RMSE[4]. For SB prediction however, the predicted value $\hat{y}_j$ is considered correct if the actual value y_j is less or equal to y_j+MAE or RMSE[5].

In general, low values of MAE or RMSE indicate a good model. However, there is no absolute criterion for a good value of MAE or RMSE as it depends on the units in which the variable is measured and on the degree of predictive accuracy, as measured in those units, which is desirable in a particular application [12]. Depending on the unit of measurement, the MAE or RMSE of the best model could be measured in hundreds, thousands or even millions. Thus, it does not makes sense to say 'the model is good or bad' because the MAE or RMSE is 'less or greater than a particular value', unless you are referring to a specific degree of accuracy that is relevant to a particular prediction application.

The recommended way to ascertain the 'goodness of fit' for such non-linear regression model is to measure the standard error of the estimate, S. This metric provides the absolute measure of the typical distance that the predicted data points fall from the regression line drawn with the true values. In other words, S is the measure of an observation made around the computed regression line. Thus, it provides an indication of the likely accuracy of predictions made with the regression line of the actual values. S is computed as:

$$S = \sqrt[2]{\frac{\sum(\hat{y} - y)^2}{n-2}} \tag{5}$$

where $\hat{y}$ represents the predicted values, y is the actual values and n is the total number of samples examined. The smaller the value of S, the less the spread and the more likely it is that any sample mean is close to the population mean. This allows for comparison between our approach and Sanders et al.'s [51]. For example, a smaller value of S for our model indicates that it is better than the *Baseline* and vice versa.

R-squared (R^2) is another 'goodness of fit' metric commonly used in regression tasks. It explains to what extent the variance of an *output variable* explains the variance of the *input variable(s)*. However, R^2 has been empirically proven to be inadequate for measuring non-linear models [54, 44, 32].

[4]A MVPA prediction must be greater or equal to a specified cut-point to be considered correct

[5]A SB prediction must be less or equal to a specified cut-point to be considered correct

Table 4
Personalised cut-points for the sample population

	Validation	Min (mg)	Max (mg)	Mean ± SD (mg)
GA	SB_{Youden}	3.14	53.91	20.87 ± 12.00
	$MVPA_{Youden}$	1.47	76.53	36.81 ± 19.82
	SB_{Se}	20.76	248.19	118.71 ± 57.80
	$MVPA_{Sp}$	202.17	976.07	454.70 ± 176.70
AG	SB_{Youden}	0.08	18.73	5.89 ± 4.60
	$MVPA_{Youden}$	0.64	37.74	17.40 ± 8.60
	SB_{Se}	0.08	179.87	16.87 ± 30.85
	$MVPA_{Sp}$	16.04	179.87	87.82 ± 49.94

Thus, only S values were used in this paper to compare performance between the *Baseline* and our approach.

3.3. Experimental Setup

We experiment with the method introduced in Section 3.2, which takes as input a set of independent characteristics about an individual, i.e., the features in Table 3, and uses them to generate *personalised* cut-points for SB and MVPA.

At first, we used ROC analysis to determine four cut-points, i.e., SB_{Youden}, $MVPA_{Youden}$, SB_{Se} and $MVPA_{Sp}$, for each of the devices, GA and AG. Instead of the collective approach by Sanders et al. [51], where a randomly counterbalanced sample of 34 individuals was used collectively to determine *generic* cut-points, we computed cut-points per individual. These *personalised* cut-points are shown in Table 4.

Using as *output features* the four individual cut-points, i.e., SB_{Youden}, $MVPA_{Youden}$, SB_{Se} and $MVPA_{Sp}$, we conducted eight separate regression experiments with the method described in Section 3.2, i.e., four experiments for each device, GA or AG. Two variations of the data described in Table 3 were used as *input features* to the experiments. Firstly, the *Basic* dataset involving only the top 7 features in Table 3; and secondly the *Enriched* dataset involving all features. The aim is to compare their performance with a view to determine the best of the two alternatives for recommending classification cut-points for SB and MVPA in practice. For example, a model trained with the *Basic* dataset allows the user, e.g., a clinician, to use easily accessible patient data such as age, gender etc., to make inference about the appropriate cut-points for SB and MVPA specific to that patient. The performance of such model is particularly interesting to compare with Sanders et al.'s [51] *generic* approach because technically, the regression model predicts cut-points without knowledge of a persons PA ability. In the presence of PA data however, a model trained with the *Enriched* dataset can be used to determine cut-points (if they lead to better performance in our experiment).

Two independent validation methods, i.e., hold-out and k-fold cross validation (CV), were used. Hold-out means splitting the dataset into a 'train' and 'test' set. We used a 50% split, stratified so that the gender distribution in the data is taken into consideration. In particular, of the 33 available data samples (male = 10, female = 23), we allocated 17 (male = 5, female = 12) for training and 16 (male = 5, female = 11) for testing. On the other hand, k-fold CV means splitting the training data into k equal size subsets, such that one of the k subsets is retained as test data, and the remaining k-1 subsets are used as training data and repeating until all k subsets have been used exactly once for testing. The k results from the folds are then combined to produce a single result. For our experiments, $k = 10$, with each fold stratified according to the gender distribution in the data.

Our intuition is that 10-Fold CV may be more suitable due to the modest data size ($n = 33$) available for this research. For example, the 50:50 hold-out sets used for training and testing may not be representative of the entire data characteristics, which would limit knowledge gain for the base classifier used. In such cases, 10-Fold CV is a good alternative method because it has the benefit of allowing all data instances to be used as test instances at least once. The results of all repetitions are then averaged to produce a combined final result. For completeness and to provide a transparent view of the findings in this experiment, the results obtained from both validation methods are reported.

As noted in Section 3.2, performance evaluation is based on both MAE and RMSE calculated as in equations 2 and 4, respectively. We also calculated S values for each model using equation 5. This allows for comparison between our approach and Sanders et al. [51].

To validate the proposed method and compare with the state-of-the-art, we investigate the research questions:

RQ1: How accurate are the *generic* cut-points of Sanders et al. [51] when evaluated per individual participant?

RQ2: Does the proposed approach outperform the state-of-the-art? If yes, to what extent?

RQ3: How does the data feature(s) contribute to the performance of our method?

The first research question (RQ1) examines the performance of the *generic* cut-points when evaluated against cut-points calculated for each study participant, rather than the general approach reported by Sanders et al. [51]. We compared the *generic* cut-points to the actual per individual to see if the results are the same as in Sanders et al. [51].

The second research question (RQ2) investigates the performance improvement of the proposed approach against the state-of-the-art. We chose Sanders et al. [51] for the following reasons. First, the experimental data was made available which allows to compare results. Second, to the best of our knowledge, this is the only state-of-the-art approach to have used standardised method post-data collection to determine raw acceleration PA cut-points specifically for older adults.

The third research question (RQ3) investigates the effects of data features on the performance of our approach. The goal is to determine how the features contribute to information gain for the base classifier, in this case *additive* regression algorithm; and also to investigate whether we can use fewer features to improve results. Recommendations would be guided by the results obtained from this question.

Table 5
Cut-points for GA and AG with associated agreement calculated according to the *generic* and the *personalised* approach

Validation		Cut-pt	Acc_G	Acc_{P1}	Acc_{P2}
GA	SB_{Youden}	≤ 20	73.1	75.0	54.5
	$MVPA_{Youden}$	≥ 32	76.2	56.3	60.6
	SB_{Se}	≤ 57	67.2	12.5	15.2
	$MVPA_{Sp}$	≥ 104	68.9	**100.0**	**100.0**
AG	SB_{Youden}	≤ 6	83.3	62.5	60.6
	$MVPA_{Youden}$	≥ 19	87.3	50.0	36.4
	SB_{Se}	≤ 15	73.2	68.8	**75.8**
	$MVPA_{Sp}$	≥ 69	80.4	68.8	66.7

Cut-pt: *generic* cut-points
Acc_G: Accuracy of the group validation reported in [51];
Acc_{P1}: Accuracy of *personalised* hold-out validation
Acc_{P2}: Accuracy of *personalised* 10-fold Cross Validation

4. Results

This section presents the results of comparison between the *Baseline* and the additive regression method proposed in Section 3.2. For clarity, we dissect the research questions RQ1, RQ2 and RQ3 in separate sub-sections.

4.1. RQ1: Examination of the accuracy of the evaluation approach in Sanders et al. [51]

Table 5 shows the *Baseline* result obtained by checking for agreement between the *generic* calibration cut-points from Sanders et al. [51] and the actual cut-points calibrated per participant. For completeness, we represent the *generic* cut-points from Sanders et al. [51] in the *Cut-pt* column. The Acc_G column represents the reported validation accuracy in Sanders et al. [51]. It is important to recall that Sanders et al.[51] obtained the Acc_G values by checking for agreement between the *generic* calibration cut-points, and a collection of samples that were not used for calibration. In other words, the ENMO values from GA and AG were aggregated for those samples not used for calibration, and then checked against the *generic* calibrated cut-points for agreement. Unfortunately, such collective validation approach is impractical in real life scenarios where samples would most likely be validated individually. Thus, we recalculated the validation results by checking for agreement between the *generic* calibration cut-points, and the actual cut-points per participant. The results are presented in the Acc_{P1} column for the hold-out validation setting and the Acc_{P2} column for the 10-fold CV setting.

The results obtained with the *personalised* approach are generally worse than those reported in Sanders et al. [51], except on two instances highlighted in **bold**, i.e., $MVPA_{Sp}$ for the GA device and SB_{Se} for the AG device. The AG SB_{Se} cut-point (≤ 15) seems reasonable for the study population. This is because the mean SB_{Se} is 16.87 mg and

Table 6
MAE, RMSE and S values used to compare the regression and *generic* [51] models in predictiong GA related cut-points

Evaluation	Cut-pt	Baseline MAE	Baseline RMSE	Baseline S	Basic MAE	Basic RMSE	Basic S	Enriched MAE	Enriched RMSE	Enriched S
Hold-out	SB_{Youden}	2.22	10.87	10.78	10.76	14.99	14.93	13.41	15.36	15.29
	SB_{Se}	62.16	85.90	85.89	46.76	59.23	59.21⋆	57.09	68.34	68.33⋆
	$MVPA_{Youden}$	8.39	24.75	24.71	16.35	19.50	19.45⋆	18.39	22.83	22.78⋆
	$MVPA_{Sp}$	392.80	436.16	436.16	207.79	270.63	270.63⋆	185.35	246.01	246.00⋆
10-Fold CV	SB_{Youden}	0.87	11.85	16.96	12.13	14.74	14.67⋆	12.91	15.60	15.53⋆
	SB_{Se}	61.71	83.96	120.56	61.82	78.36	78.35⋆	66.21	82.53	82.52⋆
	$MVPA_{Youden}$	4.81	20.10	28.83	20.24	24.57	24.52⋆	19.05	24.76	24.72⋆
	$MVPA_{Sp}$	350.70	391.49	562.24	174.16	215.09	215.09⋆	174.81	246.67	246.67⋆

⋆ Regression model better than baseline in terms of S value

standard deviation is 30.85 mg as shown in Table 4. Therefore, one could argue that the *generic* cut-point did perform well in this case, particularly with 10-fold CV with accuracy of 75.8%. However, the same does not hold for the GA $MVPA_{Sp}$ cut-point (≥ 104). Although it produced 100% accuracy with both the hold-out and the 10-fold CV, the actual $MVPA_{Sp}$ calculated for each individual participant suggests that the cut-point is extremely low. For example, the *generic* cut-point is 98.17 mg lower than the minimum actual value per individual (n = 202.17 mg) and the mean $MVPA_{Sp}$ for the study population is 454.70 mg. This is a clear indication that the *generic* cut-point is seriously under-estimated.

Apart from $MVPA_{Sp}$ for the GA device and SB_{Se} for the AG device, all other results obtained through the *personalised* validation approach produced lower accuracy values than those reported in Sanders et al. [51]. In other words, the 'one size fits all' assumption in Sanders et al. [51] is clearly impractical in real world scenario as indicated by the results in Table 5.

4.2. RQ2: Our approach vs. the state-of-the-art

Here, we compare the performance of the *generic* approach [51] with the *personalised* one proposed in this paper. For simplicity, the results are presented separately in Sections 4.2.1 for the GA device and 4.2.2 for the AG device.

4.2.1. Results for the GA Device (worn on non dominant wrist)

This section presents the comparison of results between the *Baseline* and additive regression models trained with data from the GA device. As noted in Section 3.3, the performance of the models was evaluated as an error of the predicted value using MAE and RMSE. These errors along with S values used for comparison are shown in Table 6. The results for each cut-point, i.e., SB_{Youden}, $MVPA_{Youden}$, SB_{Se} and $MVPA_{Sp}$ is presented separately with respect to the validation approach used, i.e., hold-out and 10-fold CV. Where applicable, we use '⋆' to indicate cases where the regression model is better than the *Baseline* result. This is determined by the S values, i.e., lower S value indicates superiority.

Hold-out analysis: The results obtained with hold-out validation is presented at the top part of Table 6. Using S as criterion measure, the regression models mostly performed better than the *Baseline* model. Both the *Basic* and *Enriched* regression models performed better than the *Baseline* in three out of the four cut-points when validated with the hold-out method. For example, the lower S value of 59.21 and 68.33 produced by the *Basic* and *Enriched*, respectively, mean that both models are better than the *Baseline* in predicting SB_{Se}. In other words, the distance of the predicted SB_{Se} values from the actual cut-point is shortest with the *Basic* model, followed by the *Enriched* model and farthest with the *Baseline* model. The variation between the S value of the *Baseline* and the *Basic* model is 26.68, while that of the *Enriched* model is 17.56. This is expected because the *generic* cut-point from Sanders et al. [51] did not perform particularly well in predicting SB_{Se}, in the adjusted accuracy results for GA device (i.e., Acc_{P1}) observed in Table 5. Specifically, the *generic* cut-point only predicted 12.5% of SB_{Se} correctly, when validated with *personalised* cut-points using hold-out validation. Therefore, the wide variation observed in the S values between the *Baseline* and regression models is not surprising.

The regression models were also better than the *Baseline* in predicting $MVPA_{Youden}$ and $MVPA_{Sp}$. For $MVPA_{Youden}$ cut-point prediction, the S values for *Baseline* is 24.71 while *Basic* and *Enriched* models produced 19.45 and 22.78, respectively. The *Baseline* S value is 5.27 smaller than that of *Basic* and 1.93 for *Enriched*. These values are relatively smaller than the observed differences in predicting SB_{Se}. However, this is not surprising, considering that the *Baseline* predicted 56.3% of $MVPA_{Youden}$ correctly, when we compared the *generic* cut-point from Sanders et al. [51] with *personalised* ones using hold-out validation for GA device in Table 5. Technically, higher MAE, RMSE or S indicates relative inferiority. This is the case for the *Baseline* model, which produced higher values across the three metrics than both regression models in predicting $MVPA_{Youden}$.

Much larger improvement in terms of S value was observed in $MVPA_{Sp}$ prediction, where the regression models also outperformed the *Baseline*. Interestingly, the *generic* cut-point from Sanders et al. [51] predicted 100% of

MVPA_{Sp} correctly when compared to *personalised* ones for GA device as shown in Table 5. As explained in section 4.1, there is evidence that the *generic* cut-point for MVPA_{Sp} was set too low. As such, it is likely that the *Baseline* model is not as good as it seems, in spite of achieving perfect prediction results. Indeed, this is confirmed by the lower S values produced by the regression models, showing that the predicted values have closer and better fit to the regression line than the *Baseline* model.

It is important to note that the *Baseline* model was better in predicting $\text{SB}_{\text{Youden}}$, as indicated by the higher S values produced by the regression models in Table 6. This corroborates with the results shown in Table 5, where the *generic* cut-point predicted 75.0% of the *personalised* cut-points correctly, when using hold-out validation for AG device i.e., Acc_{P1}. In fact, the *Baseline* model produced lower values across the three metrics (i.e., MAE, RMSE and S) than both regression models in predicting $\text{SB}_{\text{Youden}}$ which corroborates with the level of accuracy observed in Table 5.

10-Fold CV analysis: The results for 10-fold CV are shown in the bottom half of Table 6. Using S as criterion, the regression models performed better than the *Baseline* in all the four cut-points when validated with 10-fold CV method. The superiority of the regression models is particularly sizeable in SB_{Se} and MVPA_{Sp} results. In predicting SB_{Se} the lower S values of 78.35 and 82.52 produced by the *Basic* and *Enriched* respectively, means that both models performed better than the *Baseline* which yielded an S value of 120.56. In other words, the distance of the predicted SB_{Se} values from the actual cut-point is shortest with the *Basic* model, followed by the *Enriched* model and farthest with the *Baseline* model. It is noteworthy that the *generic* cut-point from Sanders et al. [51] did not perform particularly well in predicting SB_{Se}, when compared to *personalised* ones for GA device with 10-fold cross validation. This can be seen in the Acc_{P2} column of Table 5 for GA device, where the *generic* cut-point only predicted 15.2% of the *personalised* SB_{Se} correctly. Therefore, the lower S values produced by the regression models is not surprising.

The regression models were also better than the *Baseline* in predicting MVPA_{Sp}. Here, the S values for *Basic* and *Enriched* models are 215.09 and 246.67 respectively. These are clearly lower than the 562.24 recorded for the *Baseline* model. It is important to note that the *generic* cut-point from Sanders et al. [51] predicted 100% of MVPA_{Sp} correctly when compared to *personalised* ones for GA device as shown in Table 5. However, we know from the analysis in section 4.1, that the *generic* cut-point for MVPA_{Sp} was set too low, hence the perfect accuracy result. Indeed, this is confirmed by the higher S values produced by the *Baseline* model, showing that its predicted data points are farther away from the actual *personalised* data points on the regression line than those predicted by the *Basic* and *Enriched* models.

As noted earlier, the superiority of the regression models over the *Baseline* was evident in all four cut-points including $\text{SB}_{\text{Youden}}$ and $\text{MVPA}_{\text{Youden}}$. In predicting $\text{SB}_{\text{Youden}}$, the *Baseline* model produced an S value of 16.96 which is higher than 14.67 and 15.53 produced by the *Basic* and *Enriched* regression models. Similar results were observed for $\text{MVPA}_{\text{Youden}}$ prediction, where the S value produced by the *Baseline* model (28.83) is higher than that of the *Basic* (24.52) and *Enriched* (24.72) models by 4.31 and 4.11 respectively. Again,considering that the *generic* cut-point from Sanders et al. [51] for predicting $\text{SB}_{\text{Youden}}$ and $\text{MVPA}_{\text{Youden}}$ performed reasonably well in predicting the *personalised* cut-points as shown in the Acc_{P2} column of Table 5; the sizeable variation between S values produced by the regression and *Baseline* models indicate good improvement for our approach in the right direction.

4.2.2. Results for the AG Device (worn on left hip)

This section presents the results comparison between the *Baseline* and additive regression models trained with data from the AG device. The MAE, RMSE and S results are shown in Table 7. For clarity, the results for each cut-point, i.e., $\text{SB}_{\text{Youden}}$, $\text{MVPA}_{\text{Youden}}$, SB_{Se} and MVPA_{Sp} is presented separately with respect to the validation approach used, i.e., hold-out and 10-fold CV. Where applicable, we use '$\star$' to indicate cases where the regression model is better than the *Baseline*. This is determined by the S values i.e., lower S values indicates superiority.

Hold-out analysis: The results obtained with hold-out validation are presented at the top part of Table 7. Using S as criterion, the *Baseline* models in most cases performed better than the regression model. For example, in predicting $\text{SB}_{\text{Youden}}$, the lower S value of 3.69 produced by the *Baseline* mean that it performed better than both regression models models which yielded 4.46 for *Basic* and 3.89 for *Enriched*. In other words, the distance of the predicted SB_{Se} values from the actual cut-point is shortest with the *Baseline* model, followed by the *Enriched* model and farthest with the *Basic* model. However, the variation between the S value of the *Baseline* and the regression models are minimal, particularly the *Enriched* model where the difference is only 0.2. It is important to note that the *generic* cut-point from Sanders et al. [51] predicted $\text{SB}_{\text{Youden}}$ with 62.5% accuracy, in the hold-out validation result for AG device (i.e., Acc_{P1}) observed in Table 5. This indicates that the regression models are likely to produce accuracy values within the same region of 62.5% due to the narrow difference between their S value and that of the *Baseline* model.

Similar result was observed with MVPA_{Sp} prediction where the *Baseline* model also performed better than both regression models in terms of S value. This time, the S value for the *Baseline* model is smaller than that of *Basic* and *Enriched* models by 11.58 and 3.59 respectively. However, the the *generic* cut-point from Sanders et al. [51] predicted MVPA_{Sp} with higher accuracy of 68.8%, in the hold-out validation result for AG device (i.e., Acc_{P1}) observed in Table 5.

Mixed results were observed in predicting the other two

Table 7
MAE, RMSE and S values used to evaluate the regression and *generic* [51] models in predictiong AG related cut-points

Evaluation	Cut-pt	Baseline MAE	Baseline RMSE	Baseline S	Basic MAE	Basic RMSE	Basic S	Enriched MAE	Enriched RMSE	Enriched S
Hold-out	SB_{Youden}	0.19	3.95	3.69	3.50	4.68	4.46	3.51	4.14	3.89
	SB_{Se}	8.50	42.43	42.41	17.48	41.73	41.71⋆	18.16	43.23	43.21
	$MVPA_{Youden}$	0.59	9.51	9.40	4.43	10.40	10.30	6.81	9.35	9.24⋆
	$MVPA_{Sp}$	19.87	56.32	56.30	48.35	67.89	67.88	43.35	59.90	59.89
10-Fold CV	SB_{Youden}	0.11	4.35	6.35	4.87	6.19	6.02⋆	4.72	5.54	5.36⋆
	SB_{Se}	1.87	30.44	43.69	14.10	34.22	34.19⋆	21.91	44.66	44.64
	$MVPA_{Youden}$	1.60	8.62	12.29	9.88	11.74	11.65⋆	9.99	12.69	12.62
	$MVPA_{Sp}$	18.82	52.66	75.61	44.12	53.52	53.50⋆	46.09	58.16	59.14⋆

⋆ Regression model better than baseline in terms of S value

cut-points i.e., SB_{Se} and $MVPA_{Youden}$. For SB_{Se}, the *Basic* model produced the lowest S value (41.71), followed by 42,41 for the *Basic* model and then 43.21 for the *Enriched* model. The difference between them is marginal, so both regression models are likely to predict SB_{Se} at an accuracy level similar to the observed with the *generic* cut-point from Sanders et al. [51] (68.8%), for the AG device when validated with hold-out method as shown in Acc_{P1} column in Table 5.

Mixed result was also observed in $MVPA_{Youden}$ prediction, but this time the *Enriched* model is the most superior with S value of 9.24, followed by 9.40 for the *Basic* model and the 10.30 for the *Basic* model. The difference between them in terms of S value is also marginal - 0.90 between the *Baseline* and *Basic* models and 0.16 between the *Baseline* and *Enriched* models. As shown in Table 5,the *generic* cut-point from Sanders et al. [51] produced an average performance in predicting $MVPA_{Youden}$ with an accuracy value of 50.00% for the AG device when validated with hold-out method. Given the marginal difference in S values between all three models, it is likely that they will all produce average accuracy results.

10-Fold CV analysis: The results for 10-fold CV are shown in the bottom half of Table 7. Using S as criterion, the *Basic* regression model performed better than the *Baseline* in all the four cut-points when validated with 10-fold CV method. The *Enriched* model also performed better than the *Baseline* in predicting two out of the four cut-points i.e., SB_{Youden} and $MVPA_{Sp}$. Similar to the 10-fold cross validation result observed for GA device in Table 6, the superiority of the *Basic* regression models for AG device is also particularly sizeable in SB_{Se} and $MVPA_{Sp}$ results as shown in Table 7.

In predicting SB_{Se} the lower S value of 34.19 produced by the *Basic* model is considerably better than the *Baseline* which yielded an S value of 43.69. In other words, the distance of the predicted SB_{Se} values from the actual cut-point is shortest with the *Basic* model. This is however followed by the *Baseline* model and farthest with the *Enriched* model. It is noteworthy that the *generic* cut-point from Sanders et al. [51] performed particularly well in predicting SB_{Se}, when compared to *personalised* ones for AG device with 10-fold cross validation. This can be seen in the Acc_{P2} column of Table 5 for AG device, where the *generic* cut-point predicted 75.8% of the *personalised* SB_{Se} correctly. Therefore, the lower S values produced by the *Basic* regression model is a very good improvement.

Both regression models were better than the *Baseline* in predicting $MVPA_{Sp}$. Here, the S values for *Basic* and *Enriched* models are 53.50 and 59.14 respectively. These are clearly lower than the 75.61 recorded for the *Baseline* model. Considering that the *generic* cut-point from Sanders et al. [51] predicted 66.7% of $MVPA_{Sp}$ correctly when compared to the *personalised* cut-point values as shown in Table 5, the sizeable variation between its S value and that of the regression models can only be interpreted as a very good improvement for our approach.

As noted earlier, the superiority of the *Basic* model over the *Baseline* is evident in all four cut-points including SB_{Youden} and $MVPA_{Youden}$. In fact, both regression models performed better than the *Baseline* in predicting SB_{Youden}. As shown in Table 7, the *Baseline* model produced an S value of 6.35, which is higher than 6.02 and 5.36 produced by the *Basic* and *Enriched* regression models, respectively. Considering that the *generic* cut-point from Sanders et al. [51] for predicting SB_{Youden} performed reasonably well (60.6%) in predicting the *personalised* cut-points, as shown in the Acc_{P2} column of Table 5; the variation between S values produced by the regression and *Baseline* models indicates good improvement for our approach in the right direction.

Unfortunately, only the *Basic* regression model predicted $MVPA_{Youden}$ better than the *Baseline*, where the S value produced by the *Baseline* model (12.29) is higher than that of the *Basic* model (11.65) by 0.64 but lower than the *Enriched* model (12.62) by 0.33. This time, the *generic* cut-point from Sanders et al. [51] for predicting $MVPA_{Youden}$ did not perform particularly well (36.4%) in predicting the *personalised* cut-points, as shown in the Acc_{P2} column of Table 5. Thus, it is unlikely that the superiority of the *Basic* model over the

Baseline model would result in substantial improvement. It is still an improvement nonetheless.

4.3. RQ3: Contribution(s) of data features to our method

Based on the result analysis presented in Section 4.2, 10-Fold CV clearly led to better performance than the hold-out validation method in our experiments; particularly when applied to the *Basic* regression version of our approach. Indeed, the results validates our intuition that 10-Fold CV would yield better results than hold-out due to the modest data available for this study (n = 33). For example, the hold-out sets used for training and testing may not be representative of the entire data characteristics, which would limit knowledge gain for the base classifier used. By using 10-Fold CV however, the classifier learns from all the available data and still remains objective in its prediction. In Tables 6 and 7, the *Basic* regression was shown to perform better than its *Enriched* counterpart, when 10-Fold CV was applied. Thus, we conducted further experiments with the *Basic* regression models to measure the contribution(s) of each of the seven data features towards information gain to the base algorithm, i.e., *additive* regression. For each of the cut-points considered in this study, we re-trained the *Basic* regression model with 10-Fold CV in seven iterations; removing one of the features in each iteration and calculating how well the model fits predicted values to the regression line, i.e., the S value. The goal is to determine if fewer features could lead to better prediction.

The results are presented in Table 8. The S values obtained for the reduced feature subsets are shown in the top half of the Table. For clarity and to aid comprehension, we present in the bottom half, S values for the *Baseline* and *Basic* models, when trained with all the features and validated with 10-Fold CV. These were extracted from Tables 6 and 7. Cases where the *Basic* model trained with reduced feature subset produced better (lower) S value than with full feature subset are denoted with **bold** typeface. We also highlight (in yellow), cases where the *Basic* model trained with reduced feature subset produced worse (higher) S value than the *Baseline*.

Only one of the reduced feature subset models performed below the baseline i.e., $\text{MVPA}_{\text{Youden}}$ prediction for the AG device without the 'Systolic BP' feature. This indicates that Systolic BP is a vital feature for predicting $\text{MVPA}_{\text{Youden}}$ on the AG device, and perhaps MVPA_{Sp} as well, where the S value (58.55) also increased above the original *Basic* result (53.5). Unfortunately, the same cannot be said about predicting $\text{SB}_{\text{Youden}}$ and SB_{Se} without the 'Systolic BP' feature for AG device because of the improved performance observed.

A good number of the models performed better than the original *Basic* trained with full feature set as shown in Table 8. In particular, there is almost a perfect performance improvement across all the cut-points for both GA and AG device without the 'sex' feature. In fact, this is the case with data from AG device but fell short by one cut-point for the GA device. That said, a perfect performance improvement was observed without the 'age' feature for the GA.

It is also important to note that a lot of the results became worse than the original *Basic* model as a result of feature reduction. The performance is rarely consistent across both GA and AG devices. For example, the removal of the 'weight' feature led to reduced performance in three out of four cut-points predicted for the GA device. However, the same setting led to improvement in three out of four cut-points predicted for the AG device.

Unfortunately, none of the reduced subsets led to a perfect reduction in S value across the cut-points on both GA and AG devices. Thus, we are unable to generalise and recommend a particular subset for the purpose of improving the results of the *additive* regression algorithm when validated with 10-Fold CV.

5. Discussion

The result analysis presented in Section 4 indicates that the proposed *Basic* regression approach consistently performs better than the *Baseline* models with data from both GA and AG devices when validated with 10-Fold CV. In most cases, the proposed *Enriched* regression approach also performed better than the *Baseline* models. The discussion presented in this section will put the results into context and draw attention to a number of other factors that were not (directly) taken into consideration by the evaluation metrics used (i.e., MAE, RMSE and S) and the corresponding result analysis in Section 4.

We observed from Tables 6 and 7 that the *Baseline* models perform better than some of the regression models. Interestingly, this happens even with cut-points derived from the Youden index, which provides a compromise between sensitivity and specificity, i.e., $\text{MVPA}_{\text{Youden}}$ and $\text{SB}_{\text{Youden}}$. We suspect this is mainly due to the method in which Sanders et al [51] generated the *generic* cut-points; and partially due to factors, such as the data features, validation method employed and the device from which the experimental data was obtained.

First, Sanders et al. [51] took an a posteriori research approach, in which the cut-points were generated based on knowledge of the ENMO values taken directly from the study participants. Basically, ENMO values from a subset of the experimental data were used to generate cut-points and the remaining samples were used to evaluate accuracy. It is expected that these cut-points would perform well given that the test data belongs to the same group of participants from which the cut-points were generated. In other words, the *generic* cut-points already have direct knowledge of the output being predicted.

In our approach however, we make predictions without knowledge of the ENMO values from the study participants. Specifically, the *Basic* regression model only uses basic information about the participants as input, i.e., gender, age, weight, height, bmi, systolic and diastolic BP. Despite this, the model consistently outperformed the *Baseline* models when validated with 10-Fold CV. This is most notable in its

Table 8
Training data features and associated contribution to the S value performance of *Basic* regression model

Data subset	S value for GA				S values for AG			
	SB_{Youden}	SB_{Se}	$MVPA_{Youden}$	$MVPA_{Sp}$	SB_{Youden}	SB_{Se}	$MVPA_{Youden}$	$MVPA_{Sp}$
Features - Sex	14.97	**77.36**	**24.27**	**212.20**	**5.94**	**34.11**	**11.63**	**53.04**
Features - Age	**13.59**	**66.66**	**23.85**	**176.13**	**5.77**	34.52	13.17	54.63
Features - Weight	15.23	81.81	**22.23**	225.60	**5.41**	**33.58**	**11.41**	61.08
Features - Height	**14.18**	**66.41**	25.01	233.06	6.18	**34.12**	11.96	**53.41**
Features - BMI	15.28	80.23	**23.75**	261.52	**5.52**	**31.70**	11.70	57.45
Features - SysBP	**14.41**	**72.88**	**22.32**	238.80	**5.72**	**33.05**	12.77	58.55
Features - DiaBP	**14.18**	**74.52**	**22.84**	245.43	**5.68**	**33.32**	**9.62**	63.40
All $Feature_{Basic}$	14.67	78.35	24.52	215.09	6.02	34.19	11.65	53.5
All $Feature_{Baseline}$	16.96	120.56	28.83	562.24	6.35	43.69	12.29	75.61

prediction of SB_{Se} for the AG device, where the *generic* cut-point from Sanders et al. [51] produced its best result, i.e., 75% with 10-Fold CV in the modified accuracy representation shown in Table 5. The *Basic* regression model produced S values of 34.19 which is 9.5 lower than the 43.69 produced by the *Baseline* model. This difference in favour of the *Basic* model seems reasonably high for a model that was trained without knowledge of the ENMO values. On this note, the good results obtained with our approach show that there is potential in this line of research and further exploration which looks at optimising the base algorithm, i.e., *additive* regression, or even testing with a different algorithm may well improve the results.

We suspect that validation methods may have also contributed to the instances where our approach performed lower than the *Baseline* model. In Table 6, where the experimental data was obtained from the GA device, the *Baseline* only performed better than the regression models in predicting SB_{Youden}. More importantly, this occurred when we validated with the hold-out method. As discussed earlier in this paper, we had the intuition that 10-Fold CV would provide a more objective assessment of our approach due to the modest experimental data size ($n = 33$), because it has the benefit of allowing all data instances to be used as test instances at least once. This is in contrast to the hold-out method in which the training or testing set may not be representative of the entire data characteristics, and thus limit knowledge gain for the base classifier used. To an extent, our intuition was confirmed because both regression models performed consistently better in predicting the four cut-points than the *Baseline* model, when 10-Fold CV was applied to experimental data from the GA device.

It is possible that data features may have impacted the performance of our approach, particularly the *Enriched* model which was trained with additional features related to a person's PA characteristics. This is obvious in Table7 where the experimental data was obtained from GA device. Even when 10-Fold CV was applied, the *Enriched* model is shown to perform marginally below the *Baseline* model on two occasions but the error margin is bigger when compared to the *Basic* model. This shows that the additional features had a negative effect on our approach.

Another important factor to consider is the attachment site of the devices tested in this experiments. GA is wrist worn and AG is worn on the hip. The results of our experimentation shows that our approach performed better with data from the GA device, in comparison to its AG counterpart, which yielded mixed results with a high number of cases where our approach performed below the *Baseline* model - excluding results of 10-Fold CV with the *Basic* regression model (see Table 7). It is not very clear from the experiments why this is the case but there is evidence within the literature that wrist-worn accelerometers capture energy expenditure more accurately than hip-worn monitors [15]. Recent accelerometer studies have suggested that the wrist may be a preferable attachment site, as it allows to capture arm motions during non-ambulatory activity, such as household chores, more accurately [16, 35]. Moreover, wrist-worn accelerometers are less influenced by atypical gait patterns and walking speed variability, which are both commonly observed in older adults [31]. Older people are more likely to move their wrist than hip, so we suspect that the GA device (wrist-worn) provided a more accurate and complete set of raw acceleration signals than the AG (waist-worn). As such, the data from GA device lead to better information gain to the *input features* used for our regression models. This is particularly important because of the a priori research approach we took, in comparison to Sanders et al. [51] that took an a posteriori approach.

There is also evidence in favour of wrist-worn device over their hip-worn counterparts in terms of compliance and production of uninterrupted data [14, 34, 57]. This can be illustrated with the evolving cycles of the National Health and Nutrition Examination Survey (NHANES) in the United States of America [57]. Participants of the 2003–2004 and 2005–2006 cycles were asked to wear Actigraph 7164 on a waist belt during all non-sleeping hours for seven days. However, only about 25% of the participants provided seven days of data, and this was mostly attributed to the discomfort or inconvenience of wearing a device on the hip over time, and forgetting to put the monitor back on after taking it off at night. Fast forward to the 2011–2012 and 2013–2014 cycles for which a wrist worn Actigraph GT3X+ was used, the location change had the desired effect on compliance, as 70% of the participants provided seven days of continuous triaxial accelerometer raw-signal data at 80Hz; with the additional benefit of tracking movement during sleep. Although the experimental data used in our research was obtained during non-sleeping hours, we observed a large amount of missing values in the data from AG device (hip-worn), in comparison to GA device (wrist-worn) which had no missing value(s). We removed the rows of data with missing values during experiment and we suspect that it (i.e., data shortage) had an adverse effect on the classification tasks with data from the AG device.

Data size is another important factor that may affect the results reported in this paper. The modest dataset available for this experiment is particularly less favourable to our approach, in comparison to Sanders et al's. This is due to reasons discussed previously, where calibration and test data used in Sanders et al. [51] belongs to the same group of participants from which the calibration cut-points were obtained. A good example to illustrate this is in $MVPA_{Sp}$ prediction for the GA device, where the *generic* cut-point correctly predicted all instances (accuracy = 100%) in the modified accuracy representation shown in Table 5. The $MVPA_{Sp}$ cut-point maximises specificity over sensitivity on a Receiver Operating Curve (ROC). This happens at the point on the ROC that is capable of ruling out participants who are not engaging in MVPA, without necessarily trying to find those engaging in MVPA. Given that the test data belongs to the same participants, this cut-point is likely to result in a high but misleading accuracy value, because it was calibrated with knowledge of ENMO values from the study participants. On the other hand, our approach relies on the variation in the personal characteristics of the study participants, without knowledge of the ENMO values. Therefore, having a larger training dataset that represents a much wider variation of individual characteristics would be beneficial.

6. Conclusion

The use of accelerometer-based data in physical activity (PA) research have brought tremendous advances. Users are now able to capture, store and/or transmit large volumes of raw acceleration signal data. These data provides opportunities to characterise and represent PA better, but the opportunities are accompanied by several challenges, such as PA data analysis and interpretation. A notable challenge identified in this research is the wide variety of predictive equations available for characterising PA levels. As shown in Table 1, PA estimates derived from these equations are conceptually incompatible even though they are expressed in the same metrics. This diversity reduces the ability to make direct data comparisons between PA research studies, even when they used the same accelerometer.

Thankfully, there is growing effort in PA research to shift away from multiple independent calibration studies and move towards a consensus analytic method. This is evident in Sanders et al. [51] who opted for a post-data collection analytical process with Receiver Operating Curve (ROC) analysis. The results were promising but left some questions unattended, such as (a) the use of a single feature for PA characterisation, i.e., ENMO values derived from the accelerometer, (b) the *generic* nature of the cut-points, i.e., the one size fits all approach, and (c) the potential bias in their validation approach, because they calibrated and tested with data from the same group of individuals. We extended this research by using multiple features relating to a person's individual characteristics to predict PA cut-points. In addition, adopting a machine learning approach allowed us to personalise PA cut-points with improved performance against the state-of-the-art, particularly when validated with 10-Fold CV.

An advantage of our approach is that it can be easily replicated, thereby providing greater methodological transparency and improved comparability between different studies and accelerometer devices. Of course some logistical issues still remain, such as data availability, which is compounded by issues of compliance and attachment site for the accelerometer device, i.e., wrist, hip etc. The wrist has been highly recommended by the wider PA research community and advances in data storage, transmission, and big data computing will hopefully minimise the logistic challenges.

It may be possible to improve the performance of our approach. For example, we could apply additional optimisation techniques to the *additive* regression algorithm such as parameter tuning or replace the underlying machine learning algorithm used *Decision Stump*. In fact, there are many other simpler regression models that could be used such as the non-linear least squares [37] among others. In the future, we would consider optimisation options with the *additive* regression algorithm and compare results with other regression algorithms.

Acknowledgements

The first, third and last author have participated in this research work as part of the TYPHON Project, which has received funding from the European Unions Horizon 2020 Research and Innovation Programme under grant agreement No. 780251.

References

[1] Ainsworth, B.E., Haskell, W.L., Herrmann, S.D., Meckes, N., Bassett, D.R., Tudor-Locke, C., Greer, J.L., Vezina, J., Whitt-Glover, M.C., Leon, A.S., 2011. 2011 Compendium of Physical Activities. Medicine & Science in Sports & Exercise 43, 1575–1581. URL: https://insights.ovid.com/crossref?an=00005768-201108000-00025, doi:10.1249/MSS.0b013e31821ece12.

[2] ATS Committee on Proficiency Standards for Clinical Pulmonary Function Laboratories, 2002. ATS Statement. American Journal of Respiratory and Critical Care Medicine 166, 111–117. URL: http://www.atsjournals.org/doi/abs/10.1164/ajrccm.166.1.at1102, doi:10.1164/ajrccm.166.1.at1102.

[3] Barber, S.E., Forster, A., Birch, K.M., 2015. Levels and patterns of daily physical activity and sedentary behavior measured objectively in older care home residents in the united kingdom. Journal of Aging and Physical Activity 23, 133–143. URL: https://journals.humankinetics.com/view/journals/japa/23/1/article-p133.xml, doi:10.1123/JAPA.2013-0091.

[4] Barnett, A., van den Hoek, D., Barnett, D., Cerin, E., 2016. Measuring moderate-intensity walking in older adults using the ActiGraph accelerometer. BMC Geriatrics 16, 211. URL: http://bmcgeriatr.biomedcentral.com/articles/10.1186/s12877-016-0380-5, doi:10.1186/s12877-016-0380-5.

[5] Bing, H., Jiawei, B., Vadim, V.Z., Annemarie, K., Paolo, C., Brittney, L.M., Nancy, W.G., Tamara, B.H., Ciprian, M.C., 2014. Predicting human movement with multiple accelerometers using movelets. Medicine & Science in Sports & Exercise 46, 1859–1866. URL: https://insights.ovid.com/crossref?an=00005768-201409000-00022, doi:10.1249/MSS.0000000000000285.

[6] Blodgett, J., Theou, O., Kirkland, S., Andreou, P., Rockwood, K., 2015. The association between sedentary behaviour, moderate–vigorous physical activity and frailty in NHANES cohorts. Maturitas 80, 187–191. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378512214003636, doi:10.1016/j.maturitas.2014.11.010.

[7] Brage, S., Wedderkopp, N., Franks, P.W., Anderson, B.L., Froberg, K., 2003. Reexamination of Validity and Reliability of the CSA Monitor in Walking and Running. Medicine & Science in Sports & Exercise 35, 1447–1454. URL: https://insights.ovid.com/crossref?an=00005768-200308000-00029, doi:10.1249/01.MSS.0000079078.62035.EC.

[8] Brooks, A.G., Gunn, S.M., Withers, R.T., Gore, C.J., Plummer, J.L., 2005. Predicting Walking METs and Energy Expenditure from Speed or Accelerometry. Medicine & Science in Sports & Exercise 37, 1216–1223. URL: https://insights.ovid.com/crossref?an=00005768-200507000-00020, doi:10.1249/01.mss.0000170074.19649.0e.

[9] Byrne, N.M., Hills, A.P., Hunter, G.R., Weinsier, R.L., Schutz, Y., 2005. Metabolic equivalent: one size does not fit all. Journal of Applied Physiology 99, 1112–1119. URL: https://www.physiology.org/doi/10.1152/japplphysiol.00023.2004, doi:10.1152/japplphysiol.00023.2004.

[10] Cardinal, B., Esters, J., Cardinal, M.K., 1996. Evaluation of the Revised Physical Activity Readiness Questionnaire in older adults. Medicine & Science in Sports & Exercise 28, 468–472. URL: https://insights.ovid.com/crossref?an=00005768-199604000-00011, doi:10.1097/00005768-199604000-00011.

[11] Cardinal, B.J., Cardinal, M.K., 2000. Preparticipation Physical Activity Screening within a Racially Diverse, Older Adult Sample: Comparison of the Original and Revised Physical Activity Readiness Questionnaires. Research Quarterly for Exercise and Sport 71, 302–307. URL: http://www.tandfonline.com/doi/abs/10.1080/02701367.2000.10608910, doi:10.1080/02701367.2000.10608910.

[12] Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. Measurement Error in Nonlinear Models: A Modern Perspective. 2nd ed., Chapman and Hall/CRC, Florida.

[13] Crouter, S.E., Clowers, K.G., Bassett, D.R., 2006. A novel method for using accelerometer data to predict energy expenditure. Journal of Applied Physiology 100, 1324–1331. URL: https://www.physiology.org/doi/10.1152/japplphysiol.00818.2005, doi:10.1152/japplphysiol.00818.2005.

[14] Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M.H., White, T., van Hees, V.T., Trenell, M.I., Owen, C.G., Preece, S.J., Gillions, R., Sheard, S., Peakman, T., Brage, S., Wareham, N.J., 2017. Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. PLOS ONE 12, e0169649. URL: https://dx.plos.org/10.1371/journal.pone.0169649, doi:10.1371/journal.pone.0169649.

[15] Esliger, D.W., Rowlands, A.V., Hurst, T.L., Catt, M., Murray, P., Eston, R.G., 2011. Validation of the GENEA Accelerometer. Medicine & Science in Sports & Exercise 43, 1085–1093. URL: https://insights.ovid.com/crossref?an=00005768-201106000-00022, doi:10.1249/MSS.0b013e31820513be.

[16] Evenson, K.R., Wen, F., Herring, A.H., Di, C., LaMonte, M.J., Tinker, L.F., Lee, I.M., Rillamas-Sun, E., LaCroix, A.Z., Buchner, D.M., 2015. Calibrating physical activity intensity for hip-worn accelerometry in women age 60 to 91years: The Women's Health Initiative OPACH Calibration Study. Preventive Medicine Reports 2, 750–756. URL: https://linkinghub.elsevier.com/retrieve/pii/S2211335515001230, doi:10.1016/j.pmedr.2015.08.021.

[17] Fairclough, S.J., Noonan, R., Rowlands, A.V., Van Hees, V., Knowles, Z., Boddy, L.M., 2016. Wear Compliance and Activity in Children Wearing Wrist- and Hip-Mounted Accelerometers. Medicine & Science in Sports & Exercise 48, 245–253. URL: http://https//insights.ovid.com/crossref?an=00005768-201602000-00009, doi:10.1249/MSS.0000000000000771.

[18] Freedson, P.S., Melanson, E., Sirard, J., 1998. Calibration of the Computer Science and Applications, Inc. accelerometer. Medicine & Science in Sports & Exercise 30, 777–781. URL: https://insights.ovid.com/crossref?an=00005768-199805000-00021, doi:10.1097/00005768-199805000-00021.

[19] Friedman, J.H., 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis 38, 367–378. URL: https://linkinghub.elsevier.com/retrieve/pii/S0167947301000652, doi:10.1016/S0167-9473(01)00065-2.

[20] Frisard, M.I., Broussard, A., Davies, S.S., Roberts, L.J., Rood, J., Jonge, L.d., Fang, X., Jazwinski, S.M., Deutsch, W.A., Ravussin, E., 2007. Aging, Resting Metabolic Rate, and Oxidative Damage: Results From the Louisiana Healthy Aging Study. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 62, 752–759. URL: https://academic.oup.com/biomedgerontology/article-lookup/doi/10.1093/gerona/62.7.752, doi:10.1093/gerona/62.7.752.

[21] Hall, K.S., Howe, C.A., Rana, S.R., Martin, C.L., Morey, M.C., 2013. Mets and accelerometry of walking in older adults. Medicine & Science in Sports & Exercise 45, 574–582. URL: https://insights.ovid.com/crossref?an=00005768-201303000-00024, doi:10.1249/MSS.0b013e318276c73c.

[22] He, W., Goodkind, D., Kowal, P., 2016. An Aging World: 2015. Technical Report. U.S.Census Bureau. U.S. Government Publishing Office, Washington DC. URL: https://www.census.gov/content/dam/Census/library/publications/2016/demo/p95-16-1.pdf.

[23] van Hees, V.T., Gorzelniak, L., Dean León, E.C., Eder, M., Pias, M., Taherian, S., Ekelund, U., Renström, F., Franks, P.W., Horsch, A., Brage, S., 2013. Separating Movement and Gravity Components in an Acceleration Signal and Implications for the Assessment of Human Daily Physical Activity. PLoS ONE 8, e61691. URL: https://dx.plos.org/10.1371/journal.pone.0061691, doi:10.1371/journal.pone.0061691.

[24] Heil, D.P., Higginson, B.K., Keller, C.P., Juergens, C.A., 2003. Body size as a determinant of activity monitor output during overground walking. Journal of Exercise Psychology 6, 1 – 11.

[25] Hendelman, D., Miller, K., Baggett, C., Debold, E., Freedson, P., 2000. Validity of accelerometry for the assessment of moderate intensity physical activity in the field. Medicine & Science in Sports & Exercise 32, S442–S449. URL: https://insights.ovid.com/crossref?an=00005768-200009001-00002, doi:10.1097/00005768-200009001-00002.

[26] Hildebrand, M., Van Hees, V.T., Hansen, B.H., Ekelund, U., 2014. Age Group Comparability of Raw Accelerometer Output from Wrist- and Hip-Worn Monitors. Medicine & Science in Sports & Exercise 46, 1816–1824. URL: https://insights.ovid.com/crossref?an=00005768-201409000-00017, doi:10.1249/MSS.0000000000000289.

[27] Hills, A.P., Mokhtar, N., Byrne, N.M., 2014. Assessment of Physical Activity and Energy Expenditure: An Overview of Objective Measures. Frontiers in Nutrition 1. URL: http://journal.frontiersin.org/article/10.3389/fnut.2014.00005/abstract, doi:10.3389/fnut.2014.00005.

[28] Iba, W., Langley, P., 1992. Induction of one-level decision trees, in: Proceedings of the Ninth International Workshop on Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 233–240. URL: http://dl.acm.org/citation.cfm?id=645525.757759.

[29] Jago, R., Zakeri, I., Baranowski, T., Watson, K., 2007. Decision boundaries and receiver operating characteristic curves: New methods for determining accelerometer cutpoints. Journal of Sports Sciences 25, 937–944. URL: http://www.tandfonline.com/doi/abs/10.1080/02640410600908027, doi:10.1080/02640410600908027.

[30] Keys, A., Taylor, H.L., Grande, F., 1973. Basal metabolism and age of adult man. Metabolism 22, 579–587. URL: https://linkinghub.elsevier.com/retrieve/pii/0026049573900711, doi:10.1016/0026-0495(73)90071-1.

[31] Ko, S.u., Jerome, G.J., Simonsick, E.M., Studenski, S., Ferrucci, L., 2018. Differential Gait Patterns by History of Falls and Knee Pain Status in Healthy Older Adults: Results From the Baltimore Longitudinal Study of Aging. Journal of Aging and Physical Activity 26, 577–582. URL: https://journals.humankinetics.com/view/journals/japa/26/4/article-p577.xml, doi:10.1123/japa.2017-0225.

[32] Kvalseth, T.O., 1985. Cautionary Note about R 2. The American Statistician 39, 279. URL: https://www.jstor.org/stable/2683704?origin=crossref, doi:10.2307/2683704.

[33] Kwan, M., Woo, J., Kwok, T., 2004. The standard oxygen consumption value equivalent to one metabolic equivalent (3.5 ml/min/kg) is not appropriate for elderly people. International Journal of Food Sciences and Nutrition 55, 179–182. URL: http://www.tandfonline.com/doi/full/10.1080/09637480410001725201, doi:10.1080/09637480410001725201.

[34] Lakoski, S.G., Kozlitina, J., 2014. Ethnic Differences in Physical Activity and Metabolic Risk. Medicine & Science in Sports & Exercise 46, 1124–1132. URL: https://insights.ovid.com/crossref?an=00005768-201406000-00008, doi:10.1249/MSS.0000000000000211.

[35] Landry, G.J., Best, J.R., Liu-Ambrose, T., 2015. Measuring sleep quality in older adults: a comparison using subjective and objective methods. Frontiers in Aging Neuroscience 7, 166. URL: http://journal.frontiersin.org/Article/10.3389/fnagi.2015.00166/abstract, doi:10.3389/fnagi.2015.00166.

[36] Leenders, N.Y., Sherman, W.M., Nagaraja, H.N., Kien, C.L., 2001. Evaluation of methods to assess physical activity in free-living conditions. Medicine and Science in Sports and Exercise 33, 1233–1240. URL: https://insights.ovid.com/crossref?an=00005768-200107000-00024, doi:10.1097/00005768-200107000-00024.

[37] Levenberg, K., 1944. A method for the solution of certain non-linear problems in least squares. Quarterly of Applied Mathematics 2, 164–168. URL: http://www.jstor.org/stable/43633451.

[38] Lohne-Seiler, H., Hansen, B.H., Kolle, E., Anderssen, S.A., 2014. Accelerometer-determined physical activity and self-reported health in a population of older adults (65–85 years): a cross-sectional study. BMC Public Health 14, 284. URL: http://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-14-284, doi:10.1186/1471-2458-14-284.

[39] Luhrmann, P.M., Edelmann-Schafer, B., Neuhauser-Berthold, M., 2010. Changes in resting metabolic rate in an elderly German population: cross-sectional and longitudinal data. The journal of nutrition, health & aging 14, 232–6. URL: http://www.ncbi.nlm.nih.gov/pubmed/20191259.

[40] Mañas, A., del Pozo-Cruz, B., Guadalupe-Grau, A., Marín-Puyalto, J., Alfaro-Acha, A., Rodríguez-Mañas, L., García-García, F.J., Ara, I., 2018. Reallocating Accelerometer-Assessed Sedentary Time to Light or Moderate- to Vigorous-Intensity Physical Activity Reduces Frailty Levels in Older Adults: An Isotemporal Substitution Approach in the TSHA Study. Journal of the American Medical Directors Association 19, 185.e1–185.e6. URL: https://linkinghub.elsevier.com/retrieve/pii/S1525861017306308, doi:10.1016/j.jamda.2017.11.003.

[41] Masse, L.C., Fuemmeler, B.F., Anderson, C.B., Matthews, C.E., Trost, S.G., Cattelier, D.J., Treuth, M., 2005. Accelerometer Data Reduction: A Comparison of Four Reduction Algorithms on Select Outcome Variables. Medicine & Science in Sports & Exercise 37, S544–S554. URL: https://insights.ovid.com/crossref?an=00005768-200511001-00007, doi:10.1249/01.mss.0000185674.09066.8a.

[42] Matthews, C.E., 2005. Calibration of Accelerometer Output for Adults. Medicine & Science in Sports & Exercise 37, S512–S522. URL: https://insights.ovid.com/crossref?an=00005768-200511001-00004, doi:10.1249/01.mss.0000185659.11982.3d.

[43] Menai, M., van Hees, V.T., Elbaz, A., Kivimaki, M., Singh-Manoux, A., Sabia, S., 2017. Accelerometer assessed moderate-to-vigorous physical activity and successful ageing: results from the Whitehall II study. Scientific Reports 7, 45772. URL: http://www.nature.com/articles/srep45772, doi:10.1038/srep45772.

[44] Miaou, S.P., Lu, A., Lum, H., 1996. Pitfalls of Using R 2 To Evaluate Goodness of Fit of Accident Prediction Models. Transportation Research Record: Journal of the Transportation Research Board 1542, 6–13. URL: http://trrjournalonline.trb.org/doi/10.3141/1542-02, doi:10.3141/1542-02.

[45] Oguma, Y., Osawa, Y., Takayama, M., Abe, Y., Tanaka, S., Lee, I.M., Arai, Y., 2017. Validation of Questionnaire-Assessed Physical Activity in Comparison With Objective Measures Using Accelerometers and Physical Performance Measures Among Community-Dwelling Adults Aged ⩾85 Years in Tokyo, Japan. Journal of Physical Activity and Health 14, 245–252. URL: https://journals.humankinetics.com/view/journals/jpah/14/4/article-p245.xml, doi:10.1123/jpah.2016-0208.

[46] Perkins, N.J., Schisterman, E.F., 2006. The Inconsistency of "Optimal" Cutpoints Obtained using Two Criteria based on the Receiver Operating Characteristic Curve. American Journal of Epidemiology 163, 670–675. URL: http://academic.oup.com/aje/article/163/7/670/77813/The-Inconsistency-of-Optimal-Cutpoints-Obtained, doi:10.1093/aje/kwj063.

[47] Pober, D.M., Staudenmayer, J., Raphael, C., Freedson, P.S., 2006. Development of Novel Techniques to Classify Physical Activity Mode Using Accelerometers. Medicine & Science in Sports & Exercise 38, 1626–1634. URL: https://insights.ovid.com/crossref?an=00005768-200609000-00013, doi:10.1249/01.mss.0000227542.43669.45.

[48] Rothney, M.P., Neumann, M., Béziat, A., Chen, K.Y., 2007. An artificial neural network model of energy expenditure using nonintegrated acceleration signals. Journal of Applied Physiology 103, 1419–1427. URL: https://www.physiology.org/doi/10.1152/japplphysiol.00429.2007, doi:10.1152/japplphysiol.00429.2007.

[49] Rowland, A.V., Yates, T., Davies, M., Khunti, K., Edwardson, C.L., 2016. Raw Accelerometer Data Analysis with GGIR R-package. Medicine & Science in Sports & Exercise 48, 1935–1941. URL: http://insights.ovid.com/crossref?an=00005768-201610000-00010, doi:10.1249/MSS.0000000000000978.

[50] Rowlands, A.V., 2018. Moving Forward With Accelerometer-Assessed Physical Activity: Two Strategies to Ensure Meaningful, Interpretable, and Comparable Measures. Pediatric Exercise Science 30, 450–456. URL: https://journals.humankinetics.com/view/journals/pes/30/4/article-p450.xml, doi:10.1123/pes.2018-0201.

[51] Sanders, G.J., Boddy, L.M., Sparks, S.A., Curry, W.B., Roe, B., Kaehne, A., Fairclough, S.J., 2019. Evaluation of wrist and hip sedentary behaviour and moderate-to-vigorous physical activity raw acceleration cutpoints in older adults. Journal of Sports Sciences 37, 1270–1279. URL: https://www.tandfonline.com/doi/full/10.1080/02640414.2018.1555904, doi:10.1080/02640414.2018.1555904.

[52] Schutz, Y., Weinsier, R.L., Hunter, G.R., 2001. Assessment of Free-Living Physical Activity in Humans: An Overview of Currently Available and Proposed New Measures. Obesity Research 9, 368–379. URL: http://doi.wiley.com/10.1038/oby.2001.48, doi:10.1038/oby.2001.48.

[53] Sparling, P.B., Howard, B.J., Dunstan, D.W., Owen, N., 2015. Recommendations for physical activity in older adults. BMJ 350, h100–h100. URL: http://www.bmj.com/cgi/doi/10.1136/bmj.h100, doi:10.1136/bmj.h100.

[54] Spiess, A.N., Neumeyer, N., 2010. An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach. BMC Pharmacology 10, 6. URL: http://link.springer.com/10.1186/1471-2210-10-6, doi:10.1186/1471-2210-10-6.

[55] Swartz, A.M., Strath, S.J., Bassett, D.R., O'brein, W.L., King, G.A., Ainsworth, B.E., 2000. Estimation of energy expenditure using CSA accelerometers at hip and wrist sites. Medicine & Science in Sports & Exercise 32, S450–S456. URL: https://insights.ovid.com/crossref?an=00005768-200009001-00003, doi:10.1097/00005768-200009001-00003.

[56] Troiano, R.P., Berrigan, D., Dodd, K.W., Mâsse, L.C., Tilert, T.,

McDowell, M., 2008. Physical Activity in the United States Measured by Accelerometer. Medicine & Science in Sports & Exercise 40, 181–188. URL: https://insights.ovid.com/crossref?an=00005768-200801000-00025, doi:10.1249/mss.0b013e31815a51b3.

[57] Troiano, R.P., McClain, J.J., Brychta, R.J., Chen, K.Y., 2014. Evolution of accelerometer methods for physical activity research. British Journal of Sports Medicine 48, 1019–1023. URL: http://bjsm.bmj.com/lookup/doi/10.1136/bjsports-2014-093546, doi:10.1136/bjsports-2014-093546.

[58] Watson, K.B., Carlson, S.A., Carroll, D.D., Fulton, J.E., 2014. Comparison of accelerometer cut points to estimate physical activity in US adults. Journal of Sports Sciences 32, 660–669. URL: http://www.tandfonline.com/doi/abs/10.1080/02640414.2013.847278, doi:10.1080/02640414.2013.847278.

[59] Welk, G.J., Mcclain, J., Ainsworth, B.E., 2012. Protocols for Evaluating Equivalency of Accelerometry-Based Activity Monitors. Medicine & Science in Sports & Exercise 44, S39–S49. URL: https://insights.ovid.com/crossref?an=00005768-201201001-00006, doi:10.1249/MSS.0b013e3182399d8f.

[60] Wilms, B., Ernst, B., Thurnheer, M., Weisser, B., Schultes, B., 2014. Correction factors for the calculation of metabolic equivalents (MET) in overweight to extremely obese subjects. International Journal of Obesity 38, 1383–1387. URL: http://www.nature.com/articles/ijo201422, doi:10.1038/ijo.2014.22.

[61] Witten, I.H., Hall, M.A., Frank, E., Pal, C.J., 2017. Data Mining: Practical Machine Learning Tools and Techniques. 4th ed., Elsevier. URL: https://linkinghub.elsevier.com/retrieve/pii/C20150020718, doi:10.1016/C2015-0-02071-8.

[62] Wullems, J.A., Verschueren, S.M.P., Degens, H., Morse, C.I., Onambélé, G.L., 2017. Performance of thigh-mounted triaxial accelerometer algorithms in objective quantification of sedentary behaviour and physical activity in older adults. PLOS ONE 12, e0188215. URL: https://dx.plos.org/10.1371/journal.pone.0188215, doi:10.1371/journal.pone.0188215.

[63] Yngve, A., Nilsson, A., Sjostrom, M., Ekelund, U., 2003. Effect of Monitor Placement and of Activity Setting on the MTI Accelerometer Output. Medicine & Science in Sports & Exercise 35, 320–326. URL: https://insights.ovid.com/crossref?an=00005768-200302000-00022, doi:10.1249/01.MSS.0000048829.75758.A0.

[64] Zhu, W., Wadley, V.G., Howard, V.J., Hutto, B., Blair, S.N., Hooker, S.P., 2017. Objectively Measured Physical Activity and Cognitive Function in Older Adults. Medicine & Science in Sports & Exercise 49, 47–53. URL: http://insights.ovid.com/crossref?an=00005768-201701000-00006, doi:10.1249/MSS.0000000000001079.

CRediT authorship contribution statement

Nonso Nnamoko: Conceptualization of this study, Methodology, Final data curation, Writing - Original draft preparation. **Luis Adrián Cabrera-Diego:** Contribution to methodology approach, Proofreading. **Daniel Campbell:** Contribution to methodology approach. **George Sanders:** Data collection, Initial data curation. **Stuart J. Fairclough:** Reviewing and Editing. **Ioannis Korkontzelos:** Co-ordination, Software, Final draft preparation.