Routledge
Taylor & Francis Group

Journal of Sports Sciences

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rjsp20

A consensus method for estimating physical activity levels in adults using accelerometry

Kimberly A. Clevenger, Kelly A. Mackintosh, Melitta A. McNarry, Karin A. Pfeiffer, M. Benjamin Nelson, Joshua M. Bock, Mary T. Imboden, Leonard A. Kaminsky & Alexander H.K. Montoye

To cite this article: Kimberly A. Clevenger, Kelly A. Mackintosh, Melitta A. McNarry, Karin A. Pfeiffer, M. Benjamin Nelson, Joshua M. Bock, Mary T. Imboden, Leonard A. Kaminsky & Alexander H.K. Montoye (2022): A consensus method for estimating physical activity levels in adults using accelerometry, Journal of Sports Sciences, DOI: 10.1080/02640414.2022.2159117

To link to this article: https://doi.org/10.1080/02640414.2022.2159117

JOURNAL OF SPORTS SCIENCES
https://doi.org/10.1080/02640414.2022.2159117

Routledge
Taylor & Francis Group

PHYSICAL ACTIVITY, HEALTH AND EXERCISE

A consensus method for estimating physical activity levels in adults using accelerometry

Kimberly A. Clevenger[a], Kelly A. Mackintosh[b], Melitta A. McNarry[b], Karin A. Pfeiffer[c], M. Benjamin Nelson[d,e], Joshua M. Bock[d,f], Mary T. Imboden[d,g,h], Leonard A. Kaminsky[d,i] and Alexander H.K. Montoye[d,j]

[a]Health Behavior Research Branch, Behavioral Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, Maryland, United States; [b]Applied Sports, Technology, Exercise and Medicine Research Centre , Swansea University, Swansea, Wales, United Kingdom; [c]Department of Kinesiology, Michigan State University, East Lansing, Michigan, United States; [d]Clinical Exercise Physiology Program, Ball State University, Muncie, Indiana, United States; [e]Section on Cardiovascular Medicine, Department of Internal Medicine, Wake Forest University, Winston-Salem, North Carolina, United States; [f]Department of Cardiovascular Diseases, Mayo Clinic, Rochester, Minnesota, United States; [g]Health & Human Performance Department, George Fox University, Newberg, Oregon, United States; [h]Health Enhancement Research Organization, Raleigh, North Carolina, United States; [i]Healthy Living for Pandemic Event Protection Network, Chigaco, Illinois, United States; [j]Integrative Physiology and Health Science Department, Alma College,Alma, Michigan, United States

ABSTRACT

Identifying the best analytical approach for capturing moderate-to-vigorous physical activity (MVPA) using accelerometry is complex but inconsistent approaches employed in research and surveillance limits comparability. We illustrate the use of a consensus method that pools estimates from multiple approaches for characterising MVPA using accelerometry. Participants (n = 30) wore an accelerometer on their right hip during two laboratory visits. Ten individual classification methods estimated minutes of MVPA, including cut-point, two-regression, and machine learning approaches, using open-source count and raw inputs and several epoch lengths. Results were averaged to derive the consensus estimate. Mean MVPA ranged from 33.9–50.4 min across individual methods, but only one (38.9 min) was statistically equivalent to the criterion of direct observation (38.2 min). The consensus estimate (39.2 min) was equivalent to the criterion (even after removal of the one individual method that was equivalent to the criterion), had a smaller mean absolute error (4.2 min) compared to individual methods (4.9–12.3 min), and enabled the estimation of participant-level variance (mean standard deviation: 7.7 min). The consensus method allows for addition/removal of methods depending on data availability or field progression and may improve accuracy and comparability of device-based MVPA estimates while limiting variability due to convergence between estimates.

ARTICLE HISTORY
Received 5 July 2022
Accepted 10 December 2022

KEYWORDS
Harmonization; surveillance; measurement; raw acceleration; cut point; machine learning; devices; moderate-to-vigorous physical activity

Introduction

Physical inactivity is a leading risk factor for mortality worldwide (Carlson et al., 2018), which, if eliminated, would reduce morbidity (Lee et al., 2012), improve quality of life and productivity (Piercy et al., 2018), and increase overall life expectancy in the United States by ~9 months, similar to the elimination of smoking or obesity (Lee et al., 2012). In Europe, physical inactivity costs over €80 billion per year (Centre for Economics and Business Research, 2015), with the short- and long-term effects of the COVID-19 pandemic on inactivity levels likely to increase these costs (Tison et al., 2020). As highlighted in the World Health Organization Global Action Plan on Physical Activity (2019), surveillance is fundamental to understanding and mitigating the inactivity crisis as accurate and meaningful data on physical (in)activity can inform and evaluate interventions.

Device-based assessment for physical activity (PA) surveillance is an advantageous measurement strategy because such devices are unobtrusive and capable of measuring multiple domains of PA, such as timing, frequency, intensity, and duration and therefore provide insight into what drives, or deters, PA. However, the meaningfulness of existing device-based PA surveillance data is diminished by inconsistencies in data analysis that preclude data harmonisation. Indeed, Watson et al. (2014) found that between 6.3% and 98.3% of participants met current PA guidelines depending on how hip-worn device data were analysed. While efforts to harmonise PA surveillance data have been made (Cooper et al., 2015; Stamatakis et al., 2020), they often rely on *a priori* data collection decisions, such as the use of a specific wear location, monitor, or sampling frequency, or require individual conversion equations to facilitate inter-study comparison (Bornstein et al., 2011).

The decision regarding *how* to analyse accelerometer data, particularly the classification of time spent in moderate-to-vigorous PA (MVPA), in individual studies or surveillance systems is difficult due to the sheer number of available methods (Clevenger et al., 2022; Migueles et al., 2017; Pfeiffer et al., 2022). When selecting a classification method for adults wearing a hip-worn device, researchers could use count-based cut-points (e.g., Freedson et al., 1998), acceleration-based cut-points (e.g., Vähä-Ypyä et al., 2015), or more complex methods, such as individually-calibrated cut-points (e.g., Ozemek et al., 2013), two-regression (e.g., Crouter et al., 2010) or machine

CONTACT Kimberly A. Clevenger kimberly.clevenger@usu.edu Clevenger Health Behavior Research Branch Behavioral Research, United States

learning approaches such as artificial neural networks (e.g., Montoye et al., 2015). There are often inconsistencies in the device brand/model or sample characteristics compared to the original validation study, leading to further confusion over which approach to use for optimal accuracy in data analyses. Furthermore, the reliance on a single approach to capture PA may deter researchers from changing their analytic approach, despite new, better methods being developed, to preserve comparability with prior research.

Researchers have attempted to combine multiple methods, for example, by pooling estimates from machine learning algorithms using ensemble models (Chowdhury et al., 2017a; Chowdhury et al., 2017b; Zehra et al., 2021). Key limitations were the inclusion of only machine learning algorithms and the complexity of the pooling methods, which may limit use by other researchers. A simpler approach to combining methods was demonstrated by Troiano et al. (2008), who developed new cut-points that were a weighted average of four existing cut-points, but this is, again, limited as only one type of method and epoch length was included. In the present paper, we propose a "consensus method" that pools estimates from multiple classification approaches as a simple mean. The theoretical basis is that the pooled estimate should approach the true criterion value, despite individual classification methods each having some degree of error resulting in over- or under-estimation of MVPA. In addition to providing more stable estimates of PA outcomes, a consensus method may improve comparability with past and future research while allowing methods to evolve over time or depending on data availability. Our hypothesis for the present analysis was that the consensus approach would be equivalent to the criterion for capturing time spent in MVPA and have good stability when individual methods were added or removed from the consensus estimate.

Methods

To illustrate use of the consensus method, we conducted a secondary analysis of available data from 30 apparently healthy adults aged 18–79 years (50% females, 50% overweight or obese according to measured height and weight) who completed two laboratory visits – the first involving structured activities and the second involving semi-structured/simulated free-living activities. The study was approved by the Institutional Review Board and all participants provided written consent prior to participation. Detailed information about the methods has been previously reported (Montoye, Conger et al., 2017).

Briefly, in the first visit (~2.0–2.5 hr in length), participants completed 11 activities selected by research staff from a list of 21 possibilities, including three sedentary behaviours (e.g., lying down, writing, watching television), four household/chore activities (e.g., dusting, making the bed, sweeping), and four ambulatory/exercise activities (e.g., treadmill and overground walking, stairs, cycling) for five min each, performed in order of increasing intensity (Supplemental Table 1). Participants rested for 1–2 min between activities. The second visit incorporated simulated free-living/semi-structured activities in which participants were free to choose the order, duration, and type of activity from the same list of 21 options for 80 min, although some restrictions were imposed. For example, participants had to complete at least four activities from each category (sedentary, household/chore, ambulatory/exercise) to ensure some variety in the activity types performed during the sessions. All data, including transitions and breaks, were included in the present analysis. Both visits were included in the present study to include a range of activity types and intensities, while reflecting potential variation in how participants perform activities in free-living (Montoye, Conger et al., 2017).

A research assistant observed and recorded the activities performed and their exact start and end times during each visit. These activity types were subsequently assigned a metabolic equivalents of task (MET) value from the 2011 Compendium of Physical Activities (Ainsworth et al., (2011), with values ≥3 METs classified as MVPA. These direct observation data served as the criterion measure of time spent in MVPA, in accord with previous research (Lyden, Petruski et al., 2014).

During both visits, participants wore an ActiGraph (Pensacola, FL) GT9X triaxial accelerometer on their right hip. Devices (firmware version 1.1.0) were initialised to collect data at a sampling frequency of 60 Hz using ActiLife software (version 6.13.4), which results in comparable data to the standard 30 Hz (Brønd & Arvidsson, 2016). Data were downloaded as raw acceleration in comma-separated value (csv) format and imported to RStudio (Vienna, Austria; version 1.3.1056) using the "AGread" package (version 1.1.1). Activity counts were generated using the "agcounts" package (version 0.1.0) which uses ActiGraph's algorithm as it would be typically implemented in ActiLife software (Clevenger et al., 2022). Open-source counts were used to illustrate how this consensus method could be used with any device brand.

The outcome of interest for the present analysis was total time (min) spent in MVPA across both visits. Ten individual classification methods for estimating time spent in MVPA were used, identified according to recent systematic reviews (Clevenger et al., 2022; Migueles et al., 2017) and literature searches (Table 1). While this is not exhaustive of every available method for deriving PA intensity from hip-worn accelerometers, we chose methods ranging in complexity and data type as an illustration for how a consensus method can be implemented.

The overall consensus estimate ($MVPA_{consensus}$) was estimated for each participant as the average time spent in MVPA across all ten individual classification methods. We also tested four subsets of included classification methods, which may be necessary for other studies due to data or analytic availability. Specifically, we estimated MVPA using only count-based methods ($MVPA_{counts}$), raw acceleration-based methods ($MVPA_{raw}$), and cut-point methods ($MVPA_{cut\text{-}point}$). Additionally, due to established differences between data captured by the Computer Science and Applications (CSA) or ActiGraph 7164 monitor and newer ActiGraph models (Cain et al., 2013; Ried-Larsen et al., 2012; Whitaker et al., 2018), we tested the removal of models developed using the 7164 model ($MVPA_{no\ 7164}$).

We also generated ten "leave-one-out" consensus estimates wherein each method was excluded once. For example, $MVPA_{no\ Crouter}$ pooled estimates from the nine individual

Table 1. Classification methods for determining minutes of moderate-to-vigorous physical activity (MVPA).

Classification Method	Description of MVPA classification
Crouter et al. (2010)	Two-regression model in which the coefficient of variation in vertical axis counts·10-s^{-1} is used to determine which of two equations is used to predict METs, which are averaged over one min and classified as MVPA (≥3 METs). Implemented using the 'TwoRegression' package (version 1.0.0.9000).
Freedson et al. (1998)	Cut-point for vertical axis counts (≥1,952 counts·min^{-1})
Hibbing et al. (2018)	Two-regression model in which the coefficient of variation in ENMO·1-s^{-1} is used to determine which of two equations is used to predict METs, which are classified as MVPA (≥3 METs). Implemented using the 'TwoRegression' package (version 1.0.0.9000).
Hildebrand et al. (2014)	Cut-point for ENMO (≥69.1 m*g*) applied at a 5-s epoch
Lyden, Keadle, et al., (2014)	A combination of decision tree and artificial neural network ("Sojourn" model) implemented on vector magnitude count·s^{-1} data to predict METs, which are classified as MVPA (≥3 METs). The corrected version of the Lyden et al. (2014) method was used (Matthews et al., 2021) and implemented using the 'Sojourn' package (version 1.1.0).
Matthews et al. (2005)	Cut-point for vertical axis counts (≥760 counts·min^{-1})
Santos-Lozano et al. (2013)	Cut-point for vector magnitude counts (≥3,208 counts·min^{-1})
Sasaki et al. (2016)	Cut-point for vector magnitude counts (≥2,690 counts·min^{-1})
Troiano et al. (2008)	Cut-point for vertical axis counts (≥2,020 counts·min^{-1})
Vähä-Ypyä et al. (2015)	Cut-point for MAD (≥157.4 m*g*) applied at a 5-s epoch

ENMO: Euclidean norm minus one; MAD: mean amplitude deviation; MVPA: moderate-to-vigorous physical activity; METs: metabolic equivalents of task; m*g*: milli-*g*

methods other than Crouter et al.'s (2010) two-regression model. While these leave-one-out consensus methods are not intended for use in future research, we used them to better understand how the consensus method is affected by which individual methods are included while keeping the number of included methods consistent (n = 9).

Minutes of MVPA were compared between the criterion and the ten individual classification approaches, the overall consensus method, the four subset consensus estimates, and the ten leave-one-out consensus methods using mean absolute difference, Pearson's *r* correlation coefficient, equivalence testing, and Bland-Altman plots generated using the "blandr" package (version 0.5.1). The two one-sided tests of equivalence were conducted using the "TOSTER" package (version 0.4.0). If the 90% confidence interval around the mean difference did not overlap or exceed the equivalence bounds, the methods were considered equivalent ($p < 0.05$). Equivalence bounds were set as 10% of the mean MVPA according to the criterion (3.825 min; O'Brien, 2021). Normality was verified for all variables using histograms. All analyses were conducted in RStudio.

Results

Mean MVPA for the entire sample ranged from 33.9 to 50.4 min across the individual classification methods (Table 2, Supplemental Figure 1). MVPA estimates for each participant across the individual classification methods are illustrated in Figure 1, with a mean standard deviation across individual classification methods at the participant level of 7.7 min (range 3.0 to 20.6 min). $MVPA_{consensus}$ was 39.2 min, while the consensus estimates for subsets of included methods ranged from 38.2 ($MVPA_{cut\text{-}point}$, $MVPA_{counts}$) to 41.7 ($MVPA_{raw}$) min (Table 2). Means ranged from 38.0 ($MVPA_{no\ Hibbing}$) to 39.9 ($MVPA_{no\ Santos\text{-}Lazano}$, $MVPA_{no\ Troiano}$, $MVPA_{no\ Vähä\text{-}Ypyä}$) min for the leave-one-out consensus methods (Table 3).

Table 2. Comparison of minutes of moderate-to-vigorous physical activity (MVPA) according to the criterion, individual classification methods, and overall and subset consensus methods.

	MVPA				Absolute Difference			Equivalence Test		
Method	Mean	SD	Min	Max	Mean	SD	*r*	Bias	90% Confidence Interval	Equivalent
Criterion	38.2	6.6	25.0	49.5	-		-	-	-	-
Individual Methods										
Crouter et al. (2010)	38.9	8.6	14.0	55.0	4.9	3.9	0.69	0.7	−1.3, 2.6	Yes
Freedson et al. (1998)	35.2	11.5	3.0	50.0	5.9	6.7	0.69	−3.1	−5.7, −0.4	No
Hibbing et al. (2018)	50.4	8.5	29.2	64.1	12.3	5.8	0.69	12.1	10.2, 14.0	No
Hildebrand et al. (2014)	40.9	7.6	26.4	52.2	5.2	3.9	0.65	2.7	0.8, 4.5	No
Lyden et al. (2014)	35.6	9.7	19.2	54.9	8.1	6.1	0.31	−2.7	−5.8, 0.4	No
Matthews et al., 2005	48.3	8.9	31.0	67.0	10.3	8.3	0.41	10.0	7.3, 12.7	No
Santos-Lozano et al. (2013)	33.9	13.1	3.0	53.0	7.4	7.9	0.67	−4.3	−7.4, −1.2	No
Sasaki et al. (2016)	41.0	10.9	12.0	56.0	6.9	5.3	0.64	2.8	−0.2, 5.3	No
Troiano et al. (2008)	34.5	11.8	2.0	50.0	6.2	7.2	0.68	3.8	1.1, 6.5	No
Vähä-Ypyä et al. (2015)	33.9	7.4	18.6	46.8	6.5	3.9	0.61	4.4	2.4, 6.3	No
Consensus Methods										
All methods	39.2	8.0	18.4	51.4	4.2	3.4	0.74	1.0	−0.7, 2.7	Yes
Count methods	38.2	9.0	13.7	53.1	4.6	4.3	0.70	−0.1	−2.0, 1.9	Yes
Raw methods	41.7	7.3	28.7	53.0	5.1	3.8	0.70	3.5	1.8, 5.2	No
No 7164 methods	39.3	7.6	21.6	52.0	4.2	3.1	0.74	1.0	−0.6, 2.6	Yes
Cut-point methods	38.2	8.6	16.4	50.6	4.4	3.6	0.74	−0.0	−1.8, 1.8	Yes

MVPA: moderate-to-vigorous physical activity; SD: standard deviation; Consensus methods are the average of multiple individual methods (All: all ten individual methods; Count: the Crouter, Freedson, Lyden, Matthews, Santos-Lazano, Sasaki, and Troiano methods; Raw: the Hibbing, Hildebrand, and Vähä-Ypyä methods; No 7164: the Hibbing, Hildebrand, Lyden, Santos-Lazano, Sasaki, and Vähä-Ypyä methods; Cut-points: the Freedson, Hildebrand, Matthews, Santos-Lazano, Sasaki, Troiano, and Vähä-Ypyä methods); Confidence intervals were compared to equivalence bounds of ±3.825 min to determine equivalence at $p<0.05$.

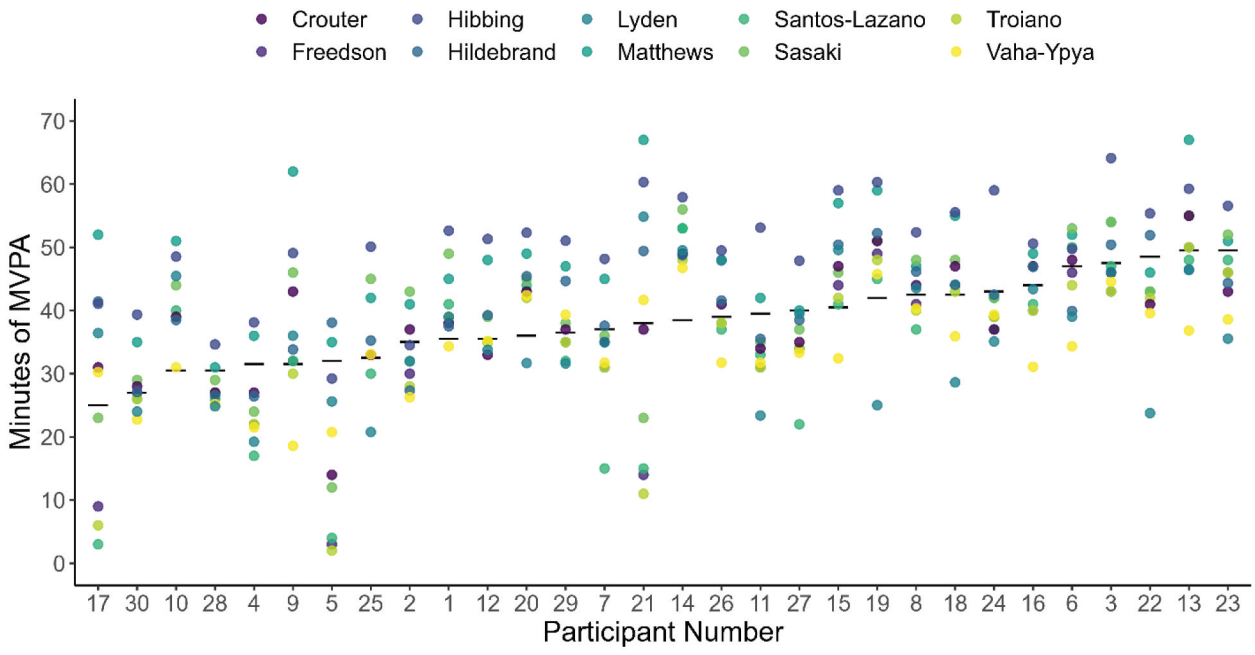

Figure 1. Minutes of moderate-to-vigorous physical activity (MVPA) for each participant across ten individual classification methods and the criterion (–).

Table 3. Comparison of minutes of moderate-to-vigorous physical activity (MVPA) between the criterion and the leave-one-out consensus methods.

	MVPA				Absolute Difference			Equivalence Test		
Method	Mean	SD	Min	Max	Mean	SD	*r*	Bias	90% Confidence Interval	Equivalent
Criterion	38.2	6.6	25.0	49.5	-		-	-	-	-
Leave-One-Out Consensus Methods										
No Crouter et al. (2010)	39.3	8.0	18.9	51.3	4.2	3.4	0.74	1.0	−0.6, 2.7	Yes
No Freedson et al. (1998)	39.7	7.8	20.1	51.6	4.4	3.3	0.74	1.4	−0.2, 3.1	Yes
No Hibbing et al. (2018)	38.0	8.1	17.2	50.5	4.1	3.6	0.74	−0.2	−1.9, 1.5	Yes
No Hildebrand et al. (2014)	39.1	8.3	16.2	51.9	4.3	3.7	0.73	0.8	−1.0, 2.6	Yes
No Lyden et al. (2014)	39.7	8.4	17.6	52.0	4.4	3.6	0.75	1.4	−0.3, 3.1	Yes
No Matthews et al., 2005	38.2	8.3	16.5	50.9	4.1	3.6	0.75	−0.0	−1.7, 1.7	Yes
No Santos-Lozano et al. (2013)	39.8	7.7	20.0	51.8	4.3	3.3	0.74	1.6	−0.0, 3.2	Yes
No Sasaki et al. (2016)	39.1	7.8	19.1	51.0	4.2	3.2	0.75	0.8	−0.8, 2.4	Yes
No Troiano et al. (2008)	39.8	7.8	20.2	51.6	4.4	3.3	0.74	1.5	−0.1, 3.2	Yes
No Vähä-Ypyä et al. (2015)	39.8	8.3	18.1	53.0	4.5	3.7	0.74	1.6	−0.2, 3.3	Yes

MVPA: moderate-to-vigorous physical activity; SD: standard deviation; Leave-one-out methods include nine of the ten individual methods, except for the one indicated (e.g., "No Crouter" includes nine methods but not Crouter et al.'s (2010) method). Confidence intervals were compared to equivalence bounds of ±3.825 min to determine equivalence at $p<0.05$.

Mean absolute differences, correlation coefficients, and equivalence testing comparing individual classification methods and consensus methods with the criterion are reported in Table 2 for individual and overall or subset consensus estimates and Table 3 for the leave-one-out consensus estimates. When compared to the criterion, $MVPA_{consensus}$ and $MVPA_{no\ 7164}$ resulted in the smallest mean absolute difference (4.2 min), which was statistically equivalent to the criterion, as were $MVPA_{counts}$, $MVPA_{cut\text{-}point}$, and all of the leave-one-out consensus methods. Mean absolute differences for individual classification methods were 4.9 to 12.3 min, with only the Crouter et al. (2010) model being equivalent to the criterion. Correlation coefficients for associations with the criterion ranged from 0.31 to 0.69 for individual classification methods, 0.70 to 0.74 for overall or subset consensus methods, and 0.73 to 0.75 for the leave-one-out consensus methods. Correlations (*r*) ranged from 0.23 to 0.99 amongst individual methods, 0.73 to 0.99 amongst consensus methods, ≥0.99 amongst the leave-one-out consensus methods, and 0.41 to 0.95 between individual and consensus methods (Supplemental Table 2).

Bland-Altman plots are presented in Supplemental Figures 2–16. All methods showed similar patterns of bias (underestimation of MVPA for the less active participants and overestimation for the more active participants), although some methods over-estimated for all participants. Consensus methods had more consistent limits of agreement and smaller biases compared to most individual methods.

Discussion

Numerous studies have demonstrated that PA estimates vary depending on how accelerometer data are analysed, including, but not limited to, cut-point selection and the use of count or raw acceleration data (Kerr et al., 2017; Loprinzi et al., 2012; Watson et al., 2014). While the identification of a single, optimal method to use for analysing accelerometer data in a specific

population would be ideal, this has not manifest in the past 40 years. Further, using a single best-practice method may still necessitate changes over time as new, better classification methods are developed, resulting in limited backward comparability with prior research. The consensus approach is a relatively simple way to pool estimates from the multitude of available methods for estimating time spent in MVPA, thereby addressing the cut-point conundrum (Trost, 2007) and/or the recent proliferation of more advanced methods (Pfeiffer et al., 2022), and improving inter-study comparisons even as methods evolve.

Whilst several studies have tried to harmonise different analytic approaches using conversion equations (Bornstein et al., 2011; Brazendale et al., 2016), it is postulated that the proposed consensus method is preferable. Specifically, instead of converting between estimates from individual classification approaches with the understanding that the MVPA estimate may be an under- or over-estimation due to inherent error for any individual method, the consensus approach works under the assumption that the mean value across classification methods will approach the true criterion value. Individual methods each have their own small bias; for example, the Hibbing et al. (2018) two-regression model consistently over-estimated MVPA. However, because some models over-estimate (Hibbing, Hildebrand, Matthews, Sasaki, Troiano, Vähä-Ypyä) and others under-estimate (Freedson, Lyden, Santos-Lazano) MVPA (Figure 1 and Supplemental Figure 1), the mean of all ten classification methods ($MVPA_{consensus}$) has reduced bias compared to individual methods. Further, new methods can theoretically be added to the consensus estimate without having to develop new conversion equations, although further testing of this feature is recommended.

An alternative to the consensus method is identifying a single "best" method of analysing accelerometer data. While not our purpose, this analysis also serves as an independent sample cross-validation of all ten approaches included in the consensus method, which is informative given that many developed methods are not independently cross-validated before being used to estimate MVPA in new populations. The Crouter et al. (2010) two-regression model performed similarly to the consensus method and was the only individual classification method that was statistically equivalent to the criterion. One potential reason individual methods do not perform well in independent samples is due to homogeneity in the sample or activities included in the original validation study. Several researchers have provided recommendations for more robust validation procedures to be employed in the development of future methods (Bassett et al., 2012; Keadle et al., 2019; Pfeiffer et al., 2022; Welk, 2005; Welk et al., 2019). However, it is noteworthy that it may still be preferable to use the consensus method for various reasons.

First, some PA surveillance systems, or indeed researchers, may not change to this single method as it would prevent backwards comparability of their findings. Using a consensus method, results from an individual method can be extracted for comparison with prior research when backward comparability is desired. Second, additional independent sample cross-validations may highlight another existing method due to inclusion of different participants, activities, settings, or devices. For example, the Sojourn model may perform better in a study employing a free-living validation protocol due to its use of bout duration to identify activity type. Future research will likely result in development of newer, more advanced, and more accurate analytic approaches. The consensus method addresses these concerns by including multiple existing methods and allowing for the inclusion of new methods. Lastly, there are situations where specific types or resolutions of data are not available, such as in earlier cycles of the National Health and Nutrition Examination Survey (Troiano et al., 2008), or the International Children's Accelerometry Database (Cooper et al., 2015). The consensus method enables researchers to use slightly different analytic methods while maintaining comparability amongst studies.

A key benefit of the consensus method is the ability to select which individual classification methods are included in the consensus estimate based on the study design and data availability. Comparability across studies is maintained because all consensus approaches theoretically average to the true criterion value and often contain overlapping methods. In the present study, whilst correlations amongst individual methods were more varied (r = 0.23 to 0.99), correlations amongst subsets of the consensus method were consistently strong ($r \geq 0.732$ to 0.99), with the weakest relationship being for the only comparison that had no overlapping methods ($MVPA_{raw}$ versus $MVPA_{counts}$; Supplemental Table 2). Further, the range in mean MVPA amongst consensus methods was five-fold smaller compared to the maximum difference amongst individual methods (3.5 versus 16.5 min). Overall, these findings provide preliminary support for improved comparability when employing consensus methods instead of individual classification methods.

While all consensus estimates were highly correlated to each other, supporting comparability across studies employing different consensus subsets, one of the consensus subsets ($MPVA_{raw}$) was not equivalent to the criterion. More research is warranted to further develop this variation on $MVPA_{consensus}$. It is pertinent to note that the consensus method works on the assumption that individual methods over- and under-estimate MVPA, and therefore pooling results approaches the criterion value, and thus will not work if there is consistent bias across methods (i.e., error is not random). All three methods in $MVPA_{raw}$ over-estimated MVPA, resulting in over-estimation compared to the criterion. Inclusion of more than three individual methods in $MVPA_{raw}$ might reduce bias for this consensus method, which will be useful to researchers using non-ActiGraph devices. Future research in other populations may also reveal systematic error of individual methods, so the validity of consensus methods should be tested prior to application in new populations.

The leave-one-out analyses allowed us to further understand how the combination of included methods may impact the consensus estimate while keeping the number of included methods consistent. All the leave-one-out consensus methods were equivalent to the criterion and highly correlated with each other, irrespective of which methods were included. Even removal of the most accurate individual classification method resulted in a consensus estimate that was equivalent to the criterion, despite it being comprised of nine methods which

were not equivalent to the criterion individually. Overall, the strong correlations amongst consensus methods and the leave-one-out analyses support that which methods are included in the consensus estimate should minimally affect MVPA estimates, but researchers should confirm how the addition or replacement of included classification methods affects the consensus estimate as this was not fully tested in the present study.

We selected the included methods based on prior use and to include a variety of data types, epoch lengths, and analytic complexity. While further research is needed, particularly in free-living settings, with different samples and devices, the validity demonstrated in the present analysis provides preliminary support for use of the consensus method described in this paper in a general adult population wearing an ActiGraph device at the right hip. As more of the individual classification methods are cross-validated and better classification methods are developed, researchers may identify the optimal number or combination of included methods and establish standards for inclusion or exclusion of individual classification methods in the consensus approach. For example, minimum criteria for inclusion of a particular method could be that, in addition to being made openly available, a model should be within some degree of accuracy compared to a criterion, have been validated in a sample of a certain size or heterogeneity, have been independently cross-validated, and developed using semi-structured or unstructured activities. To reduce complexity, it could be determined that the consensus method should include a maximum number of individual classification methods (e.g., ten) and therefore the addition of new models should only be considered when they out-perform prior methods. Further research should ascertain the trade-off between increasing the number of included methods, precision, and difficulty of implementation.

Another benefit of the consensus method is that it allows for estimation of variance at the participant level. Similar to work in nutritional epidemiology (National Cancer Institute, 2022), this estimated variance could be used in more advanced modelling approaches to adjust for each participant's potential measurement error. Further research can also establish consensus methods for devices at different wear locations (e.g., wrist) or in other populations (e.g., children, older adults). Consensus methods may also be developed for other behaviours, such as sleep, while epoch-level consensus methods could be used for identification of non-wear time. Additionally, future research could identify if methods other than a simple mean would be preferable, such as an average weighted by the quality or size of the initial validation protocol or the model's accuracy compared to a criterion. This work could be informed by prior research on meta-analyses or ensemble models (Chowdhury et al., 2017a; Chowdhury et al., 2017b; Zehra et al., 2021). Finally, consensus methods could be used to pool estimates from multiple devices, wear locations, or with different outcomes (METs, activity intensity, activity type).

This study and the consensus method are not without limitations. While participants completed a semi-structured, simulated free-living protocol, we did not include true free-living measurements commensurate with Lyden et al. (2014). The accuracy of individual or consensus methods should be verified using free-living observations, potentially using direct observation as the criterion, given that habitual behaviours may involve different activity types or movement patterns. Longer validation protocols could also examine whether consensus methods are able to capture other characteristics of PA, such as frequency or timing, and include more detailed analysis regarding the intensity level (i.e., disaggregating moderate- and vigorous- intensity activity). While the consensus method is not computationally difficult, some may find it cumbersome to use multiple epoch lengths and methods to analyse their data. Nonetheless, open-source code or software packages could be developed to streamline this process.

Importantly, there are available classification methods we did not include in the consensus estimate. For example, cut-points developed on the CSA or ActiGraph 7164 accelerometer are not ideal due to known differences in data recorded by newer ActiGraph models (Cain et al., 2013; Ried-Larsen et al., 2012; Whitaker et al., 2018). However, some methods originally developed on older models that have been continuously used over the years in newer ActiGraph models, such as the Freedson et al. (1998) cut-points, were included in our consensus method to provide some degree of backwards comparability. Some models were duplicative, such as Lyden, Keadle et al.'s (2014) vertical axis versus vector magnitude Sojourn model, whereas other models are available that were developed for the purpose of answering a specific question and not necessarily intended for use in other research, such as machine learning models used to test different modelling approaches or feature sets (Montoye, Begum et al., 2017). We reiterate that the premise of the consensus approach is not to include all available methods to assess PA intensity (or whatever the outcome of interest), but rather that combining outputs from several individual methods should have higher accuracy and better stability than individual methods and improve comparability amongst studies.

In conclusion, the consensus method is a simple approach which may improve inter-study comparability while still enabling flexibility in analytical approach based on data availability or future research developments. Indeed, our work demonstrates that the mean value (MVPA estimate) is only minimally affected by adding or removing individual classification methods from the consensus approach, and therefore enables changes in classification methods as the field of PA measurement progresses while maintaining inter-study comparability. Additional research on consensus methods in other populations or outcomes is warranted.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Ball State University.

ORCID

Kimberly A. Clevenger http://orcid.org/0000-0003-2993-3587

References

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. (20112011). Compendium of Physical Activities: A second update of codes and MET values. *Medicine and Science in Sports and Exercise*, *43* (8), 1575–1581. https://doi.org/10.1249/MSS.0b013e31821ece12

Azeem, N., H, S., & Farhan, M. (2021). Human activity recognition through ensemble learning of multiple convolutional neural networks. *2021 55th Annual Conference on Information Sciences and Systems*. https://doi.org/10.1109/CISS50987.2021.9400290

Bassett, D. R., Jr, Rowlands, A. V., & Trost, S. G. (2012). Calibration and validation of wearable monitors. *Medicine and Science in Sports and Exercise*, *44*(1 Suppl 1), S32. https://doi.org/10.1249/MSS.0b013e3182399cf7

Bornstein, D. B., Beets, M. W., Byun, W., Welk, G., Bottai, M., Dowda, M., & Pate, R. (2011). Equating accelerometer estimates of moderate-to-vigorous physical activity: In search of the Rosetta Stone. *Journal of Science and Medicine in Sport*, *14*(5), 404–410. https://doi.org/10.1016/j.jsams.2011.03.013

Brazendale, K., Beets, M. W., Bornstein, D. B., Moore, J. B., Pate, R. R., Weaver, R. G., Falck, R. S., Chandler, J. L., Andersen, L. B., & Anderssen, S. A. (2016). Equating accelerometer estimates among youth: The Rosetta Stone 2. *Journal of Science and Medicine in Sport*, *19*(3), 242–249. https://doi.org/10.1016/j.jsams.2015.02.006

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

Cain, K. L., Conway, T. L., Adams, M. A., Husak, L. E., & Sallis, J. F. (2013). Comparison of older and newer generations of ActiGraph accelerometers with the normal filter and the low frequency extension. *International Journal of Behavioral Nutrition and Physical Activity*, *10*(1), 1–6. https://doi.org/10.1186/1479-5868-10-51

Carlson, S. A., Adams, E. K., Yang, Z., & Fulton, J. E. (2018). Peer reviewed: Percentage of deaths associated with inadequate physical activity in the United States. *Preventing Chronic Disease*, *15*, 170354. https://doi.org/10.5888/pcd18.170354

Centre for Economics and Business Research. (2015). *The Economic Cost of Physical Inactivity in Europe*.

Chowdhury, A. K., Tjondronegoro, D., Chandran, V., & Trost, S. (2017). Ensemble methods for classification of physical activities from wrist accelerometry. *Medicine and Science in Sports and Exercise*, *49*(9), 1965–1973. https://doi.org/10.1249/MSS.0000000000001291

Chowdhury, A. K., Tjondronegoro, D., Chandran, V., & Trost, S. G. (2017). Physical activity recognition using posterior-adapted class-based fusion of multiaccelerometer data. *Journal of Biomedical and Health Informatics*, *22*(3), 678–685.

Clevenger, K. A., Montoye, A. H., Van Camp, C. A., Strath, S. J., & Pfeiffer, K. A. (2022). Methods for estimating physical activity and energy expenditure using raw accelerometry data or novel analytical approaches: A repository, framework, and reporting guidelines. *Physiological Measurement*, *43*(9), 9. https://doi.org/10.1088/1361-6579/ac89c9

Cooper, A. R., Goodman, A., Page, A. S., Sherar, L. B., Esliger, D. W., van Sluijs, E. M., Andersen, L. B., Anderssen, S., Cardon, G., Davey, R., Froberg, K., Hallal, P., Janz, K. F., Kordas, K., Kreimler, S., Pate, R. R., Puder, J. J., Reilly, J. J., Salmon, J., Ekelund, U. (2015). Objectively measured physical activity and sedentary time in youth: The International children's accelerometry database (ICAD). *International Journal of Behavioral Nutrition and Physical Activity*, *12*(1), 1–10. https://doi.org/10.1186/s12966-015-0274-5

Crouter, S. E., Kuffel, E., Haas, J. D., Frongillo, E. A., & Bassett, D. R., Jr. (2010). A refined 2-regression model for the actigraph accelerometer. *Medicine and Science in Sports and Exercise*, *42*(5), 1029. https://doi.org/10.1249/MSS.0b013e3181c37458

Freedson, P. S., Melanson, E., & Sirard, J. (1998). Calibration of the Computer Science and Applications, Inc. accelerometer. *Medicine and Science in Sports and Exercise*, *30*(5), 777–781. https://doi.org/10.1097/00005768-199805000-00021

Hibbing, P. R., Lamunion, S. R., Kaplan, A. S., & Crouter, S. E. (2018). Estimating energy expenditure with actigraph gt9x inertial measurement unit. *Medicine and Science in Sports and Exercise*, 50(5), 1093–1102. https://doi.org/10.1249/MSS.0000000000001532

Hildebrand, M., Vt, V. A. N. H., Hansen, B. H., & Ekelund, U. (2014). Age group comparability of raw accelerometer output from wrist- and Hip-worn monitors. *Medicine and Science in Sports and Exercise*, *46*(9), 1816–1824. https://doi.org/10.1249/mss.0000000000000289

Keadle, S. K., Lyden, K. A., Strath, S. J., Staudenmayer, J. W., & Freedson, P. S. (2019). A framework to evaluate devices that assess physical behavior. *Exercise and Sport Sciences Reviews*, *47*(4), 206–214. https://doi.org/10.1249/JES.0000000000000206

Kerr, J., Marinac, C. R., Ellis, K., Godbole, S., Hipp, A., Glanz, K., Mitchell, J., Laden, F., James, P., & Berrigan, D. (2017). Comparison of accelerometry methods for estimating physical activity. *Medicine and Science in Sports and Exercise*, *49*(3), 617. https://doi.org/10.1249/MSS.0000000000001124

Lee, I.-M., Shiroma, E. J., Lobelo, F., Puska, P., Blair, S. N., & Katzmarzyk, P. T. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy. *The Lancet*, *380*(9838), 219–229. https://doi.org/10.1016/S0140-6736(12)61031-9

Loprinzi, P. D., Lee, H., Cardinal, B. J., Crespo, C. J., Andersen, R. E., & Smit, E. (2012). The relationship of actigraph accelerometer cut-points for estimating physical activity with selected health outcomes: Results from NHANES 2003-06. *Research Quarterly for Exercise and Sport*, *83*(3), 422–430. https://doi.org/10.1080/02701367.2012.10599877

Lyden, K., Keadle, S. K., Staudenmayer, J., & Freedson, P. S. (2014). A method to estimate free-living active and sedentary behavior from an accelerometer. *Medicine and Science in Sports and Exercise*, *46*(2), 386.

Lyden, K., Petruski, N., Mix, S., Staudenmayer, J., & Freedson, P. (2014). Direct observation is a valid criterion for estimating physical activity and sedentary behavior. *Journal of Physical Activity & Health*, *11*(4), 860–863.

Matthews, C. E. (2005). Calibration of accelerometer output for adults. *Medicine and Science in Sports and Exercise*, *37*(11 Suppl), S512–522. https://doi.org/10.1249/01.mss.0000185659.11982.3d

Matthews, C. E., Keadle, S. K., Berrigan, D., Lyden, K., & Troiano, R. P. (2021). Influence of accelerometer calibration approach on moderate-vigorous physical activity estimates for adults—corrigendum. *medicine and science in sports and exercise*. 53. 9. https://doi.org/10.1249/MSS.0000000000002669

Migueles, J. H., Cadenas-Sanchez, C., Ekelund, U., Nyström, C. D., Mora-Gonzalez, J., Löf, M., Labayen, I., Ruiz, J. R., & Ortega, F. B. (2017). Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations. *Sports Medicine*, *47*(9), 1821–1845. https://doi.org/10.1007/s40279-017-0716-0.

Montoye, A. H., Begum, M., Henning, Z., & Pfeiffer, K. A. (2017). Comparison of linear and non-linear models for predicting energy expenditure from raw accelerometer data. *Physiological Measurement*, *38*(2), 343.

Montoye, A. H., Conger, S. A., Connolly, C. P., Imboden, M. T., Nelson, M. B., Bock, J. M., & Kaminsky, L. A. (2017). Validation of accelerometer-based energy expenditure prediction models in structured and simulated free-living settings. *Measurement in Physical Education and Exercise Science*, *21*(4), 223–234. https://doi.org/10.1080/1091367X.2017.1337638

Montoye, A. H., Mudd, L. M., Biswas, S., & Pfeiffer, K. A. (2015). Energy expenditure prediction using raw accelerometer data in simulated free living. *Medicine and Science in Sports and Exercise*, *47*(8), 1735–1746. https://doi.org/10.1249/mss.0000000000000597

National Cancer Institute. (2022). *Measurement Error in Nutritional Epidemiology*. Retrieved March 31 from https://prevention.cancer.gov/research-groups/biometry/measurement-error-impact/measurement-error-0#:~:text=The%20NCI%20method%20may%20be%20applied%20to%20reduce,intake%20through%20the%20NCI%20model%20for%20measurement%20error

O'Brien, M. W. (2021). Implications and recommendations for equivalence testing in measures of movement behaviors: a scoping review. *Journal for the Measurement of Physical Behaviour*, *4*(4), 353–362. https://doi.org/10.1123/jmpb.2021-0021

Ozemek, C., Cochran, H. L., Strath, S. J., Byun, W., & Kaminsky, L. A. (2013). Estimating relative intensity using individualized accelerometer cut-points: The importance of fitness level. *BMC Medical Research Methodology*, *13*(1), 1–7. https://doi.org/10.1186/1471-2288-13-53

Pfeiffer, K. A., Clevenger, K. A., Kaplan, A., Van Camp, C. A., Strath, S. J., & Montoye, A. H. (2022). Accessibility and use of novel methods for predicting physical activity and energy expenditure using accelerometry: A scoping review. *Physiological Measurement*, *43*(9), 09TR01. https://doi.org/10.1088/1361-6579/ac89ca

Piercy, K. L., Troiano, R. P., Ballard, R. M., Carlson, S. A., Fulton, J. E., Galuska, D. A., George, S. M., & Olson, R. D. J. J. (2018). The physical activity guidelines for Americans. *Jama*, *320*(19), 2020–2028. https://doi.org/10.1001/jama.2018.14854

Ried-Larsen, M., Brønd, J. C., Brage, S., Hansen, B. H., Grydeland, M., Andersen, L. B., & Møller, N. C. (2012). Mechanical and free living comparisons of four generations of the Actigraph activity monitor. *International Journal of Behavioral Nutrition and Physical Activity*, *9*(1), 1–10. https://doi.org/10.1186/1479-5868-9-113

Santos-Lozano, A., Santin-Medeiros, F., Cardon, G., Torres-Luque, G., Bailon, R., Bergmeir, C., Ruiz, J. R., Lucia, A., & Garatachea, N. (2013). Actigraph GT3X: Validation and determination of physical activity intensity cut points. *International Journal of Sports Medicine*, *34*(11), 975–982. https://doi.org/10.1055/s-0033-1337945

Sasaki, J. E., da Silva, K. S., da Costa, B. G. G., & John, D. (2016). Measurement of physical activity using accelerometers. In Luiselli James K & Fischer Aaron (Eds.), *Computer-assisted and web-based innovations in psychology, special education, and health* (pp. 33–60). Elsevier.

Stamatakis, E., Koster, A., Hamer, M., Rangul, V., Lee, I.-M., Bauman, A. E., Atkin, A. J., Aadahl, M., Matthews, C. E., & Mork, P. J. (2020). *Emerging collaborative research platforms for the next generation of physical activity, sleep and exercise medicine guidelines: The Prospective Physical Activity, Sitting, and Sleep consortium (ProPASS)*. BMJ Publishing Group Ltd and British Association of Sport and Exercise Medicine.

Tison, G. H., Avram, R., Kuhar, P., Abreau, S., Marcus, G. M., Pletcher, M. J., & Olgin, J. E. (2020). Worldwide effect of COVID-19 on physical activity: A descriptive study. *Annals of Internal Medicine*, *173*(9), 767–770. https://doi.org/10.7326/M20-2665

Troiano, R. P., Berrigan, D., Dodd, K. W., Masse, L. C., Tilert, T., & McDowell, M. (2008). Physical activity in the United States measured by accelerometer. *Medicine and Science in Sports and Exercise*, *40*(1), 181. https://doi.org/10.1249/mss.0b013e31815a51b3

Trost, S. G. (2007). State of the art reviews: Measurement of physical activity in children and adolescents. *American Journal of Lifestyle Medicine*, *1*(4), 299–314. https://doi.org/10.1177/1559827607301686

Vähä-Ypyä, H., Vasankari, T., Husu, P., Suni, J., & Sievänen, H. (2015). A universal, accurate intensity-based classification of different physical activities using raw data of accelerometer. *Clinical Physiology and Functional Imaging*, *35*(1), 64–70. https://doi.org/10.1111/cpf.12127

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*(7), 660–669. https://doi.org/10.1080/02640414.2013.847278

Welk, G. J. (2005). Principles of design and analyses for the calibration of accelerometry-based activity monitors. *Medicine and Science in Sports and Exercise*, *37*(11 Suppl), S501–11. https://doi.org/10.1249/01.mss.0000185660.38335.de

Welk, G. J., Bai, Y., Lee, J. M., Godino, J. O. B., Saint-Maurice, P. F., & Carr, L. (2019). Standardizing analytic methods and reporting in activity monitor validation studies. *Medicine and Science in Sports and Exercise*, *51*(8), 1767. https://doi.org/10.1249/MSS.0000000000001966

Whitaker, K. M., Gabriel, K. P., Jacobs, D. R., Jr, Sidney, S., & Sternfeld, B. (2018). Comparison of two generations of ActiGraph accelerometers: The CARDIA study. *Medicine and Science in Sports and Exercise*, *50*(6), 1333. https://doi.org/10.1249/MSS.0000000000001568

World Health Organization. (2019). *Global action plan on physical activity 2018-2030: More active people for a healthier world*. *World Health Organization*.

Zehra N., Azeem S. H., & Farhan M. Human Activity Recognition Through Ensemble Learning of Multiple Convolutional Neural Networks,2021 55th Annual Conference on Information Sciences and Systems (CISS), 2021, pp. 1–5. https://doi.org/10.1109/CISS50987.2021.9400290