A Consensus Method for Estimating Moderate-to-Vigorous Physical Activity Levels in Adults Using Wrist-Worn Accelerometers

Date of Submission: March 11, 2024

Abstract

Inconsistency in the calculation of time spent in moderate-to-vigorous physical activity (MVPA) limits inter-study comparability and interpretation of surveillance data. This study assesses whether combining multiple individual methods results in a more accurate estimate of MVPA, while considering the influence of device brand and wear location. Participants (n=30, age=49.2 ± 19.5 y) wore two accelerometers (GENEActiv, ActiGraph) on each wrist during two laboratory visits. Individual classification methods (11 for left wrist, 8 for right wrist) estimated minutes of MVPA using three approaches (cut-point, two-regression, machine learning), two types of input (count and raw), and five epoch lengths (1, 5, 15, 30, 60 s). The consensus estimate was calculated as the mean or median (due to skew) across all individual estimates. No individual or consensus estimates were statistically equivalent to direct observation (mean 38.2 min), with 81-95% of individual methods over-estimating MVPA. The best-performing individual methods were raw acceleration cut-points, with a bias of -3.2 to 2.4 min across devices and wrists. Correlation coefficients between individual methods and the criterion were 0.35-0.71 for the left and 0.12-0.67 for the right wrist, compared to 0.65-0.70 and 0.58-0.66 for consensus methods, respectively. Correlations between device brands were 0.23-0.99 for individual methods and 0.70-0.86 for consensus methods, whilst correlations between locations were 0.55-0.86 and 0.73-0.87, respectively. Better methods are required for estimating MVPA from wrist-worn accelerometers given the consistent over-estimation of MVPA observed. Whilst a consensus method for wrist-worn data was not able to fully resolve these issues, it improves inter-wrist or -brand comparability.

Keywords: harmonisation, surveillance, measurement, raw acceleration, cut-points, machine learning

Introduction

Moderate-to-vigorous physical activity (MVPA) is recommended for all adults due to its well-known association with cardiometabolic and psychosocial health and reduced risk of non-communicable disease and premature death (Ding et al., 2016; Kraus et al., 2019; McTiernan et al., 2019; Piercy & Troiano, 2018; Saint-Maurice et al., 2022). Accelerometers are useful tools for measuring MVPA in a variety of settings and populations because they are small and unobtrusive, not subject to participant recall bias, and can be used across languages and cultures (Pedišić & Bauman, 2015). Whilst these devices have been traditionally worn on the hip, a shift towards wrist-worn placements has occurred due to improved wear compliance and a focus on 24-h movement behaviours (Fairclough et al., 2016; Troiano et al., 2014). Indeed, national surveillance efforts in the United States and United Kingdom, for example, now utilise wrist-worn devices (Belcher et al., 2021; Doherty et al., 2017).

Many options are available to those interested in estimating MVPA from wrist-worn accelerometers, including using count- or raw acceleration-based cut-points (Dillon et al., 2016; Esliger et al., 2011; Hildebrand et al., 2014; Kwan et al., 2020; Lee & Tse, 2019; Montoye et al., 2020; Neil-Sztramko et al., 2017; Rhudy et al., 2020), two-regression models (Hibbing et al., 2018), or machine-learning algorithms (Montoye et al., 2017; Staudenmayer et al., 2015). A recent review found 67 methods for transforming wrist accelerometer data into physical activity intensity or energy expenditure solely using raw acceleration and/or machine-learning methods (Pfeiffer et al., 2022). While many of these methods were created for specific populations (e.g., children, older adults), the large number of options for a given population makes it difficult for researchers to know which method to use or is indeed optimal. Use of different processing methods is problematic because it limits our ability to compare across studies or surveillance systems.

Identifying the most accurate method is difficult because variations in sample characteristics, activities completed, setting, and how results were statistically compared to the criterion differ amongst validation studies. Further, unlike the hip where wear on the right side of the body was standard, there is no such agreement as to which wrist the device should be worn on (e.g. non-dominant wrist, right wrist) (Liu et al., 2021). Few methods have undergone independent sample cross-validation (Farrahi et al., 2019; Pfeiffer et al., 2022), which is critical to ascertain the ecological or external validity of a method and to understand how it will perform in a new setting and/or with new participants who may perform activities in unique ways (Clevenger, Montoye, et al., 2022; Montoye et al., 2018). The limited research cross-validating existing methods demonstrates that models perform worse in an independent sample than the original validation study (Ellingson et al., 2017; Montoye et al., 2018). Conducting independent sample cross-validation of multiple methods at the same time is particularly useful so that methods can be directly compared without added variability due to participant characteristics or data collection and processing decisions. Additional cross-validation research is needed to inform the optimal approach to analysing wrist-worn accelerometer data.

An added complexity is that existing methods are typically developed using a single accelerometer brand, or generation of device, and many specifically utilise the ActiGraph count metric as an input. Until recently, ActiGraph counts were calculated using a proprietary process, limiting use of these methods to studies using ActiGraph devices. Now that the count algorithm has been made open source (Neishabouri et al., 2022), researchers can theoretically use methods developed using ActiGraph count data with other devices, such as the GENEActiv. Whilst prior research has compared raw acceleration data between GENEActiv and ActiGraph devices (Rowlands et al., 2017), the validity of using methods developed with ActiGraph on GENEActiv data input has not been assessed.

Together, the large number of available methods, and the lack of independent sample cross-validation studies directly comparing these methods – particularly as researchers may use different device brands or wear locations – hinders our ability to use accelerometry as an accurate physical activity measurement tool. While it is important to accurately measure MVPA in individual studies or surveillance systems (e.g., to assess the impact of physical activity-promoting interventions, monitor trends over time), lack of agreement as to how to analyse wrist accelerometer data also limits comparability across studies or surveillance systems. Attempts have been made to improve comparability across studies. For example, the Prospective Physical Activity, Sitting, and Sleep consortium (ProPASS) has generated standard operating procedures to harmonize data collection across cohorts as well as methodologies to harmonize data after collection is completed. Alternatively, the monitor-independent movement summary (MIMS) unit was created to account for differences in parameters such as device sampling frequency or dynamic range, which should improve comparability in data collected across different device types and initialization parameters. Other approaches to harmonization include the use of conversion equations (Brazendale et al., 2016), ensemble models to pool estimates across machine-learning algorithms (Chowdhury et al., 2017), or pooled cut-points prior to application (Troiano et al., 2008).

Clevenger et al. (2022) recently proposed another solution to the challenge of harmonizing accelerometer analyses across studies. When comparing processing techniques for hip-worn accelerometer data, they found that most individual methods did not accurately predict MVPA (Clevenger, Mackintosh, et al., 2022). Given that individual methods both over- and under-estimate MVPA, it was postulated that the average across methods (the "consensus") may approach the true value (Clevenger, Mackintosh, et al., 2022). Indeed, it was demonstrated that 10 individual classification methods, including cut-points, two-regression models, and machine-learning algorithms, had mean absolute errors ranging from 4.9 to 12.3

min compared to the criterion of direct observation in adults wearing a hip-worn ActiGraph. Averaging estimates from these 10 individual methods resulted in reduction of mean absolute error to 4.2 min. The consensus method also had improved comparability with individual methods, indicating it may help resolve the issue of poor comparability across studies employing different processing methods. The consensus approach is unique in that it allows for inclusion of a variety of model types, epoch lengths, and data inputs, maximising its flexibility and application. However, the utility of such a consensus method needs to be verified at other wear locations.

The purpose of the present analysis was to assess the criterion validity of a consensus method for estimating MVPA using wrist-worn accelerometers. We hypothesised that, akin to the hip-consensus method, the wrist-consensus method would be equivalent to the criterion for capturing time spent in MVPA. This study also provides an independent sample cross-validation of the included individual methods, including the application of methods developed with ActiGraph devices to data collected using GENEActiv data (and vice-versa). Finally, we compare the consensus and individual methods across device brands (ActiGraph and GENEActiv) and wear locations (dominant and non-dominant wrist). We hypothesized that the consensus method would demonstrate improved comparability across device brands and wear locations compared to individual methods.

Methods

Data Collection

The same data used in developing the hip-consensus method were used for the present analysis (Montoye et al., 2017). Briefly, the Institutional Review Board approved the study protocol, after which 30 adults 18-79 years of age (49.2 ± 19.5y; 50% female) provided written informed consent prior to participation in structured and semi-structured laboratory visits.

Participants were apparently healthy (no known disease or disability) and did not need a physician's clearance for participation in exercise according to a Physical Activity Readiness Questionnaire (PAR-Q). Recruitment was stratified by age (18-40 y, 41-60 y, 61-80 y). Body Mass Index (BMI) was 26.0 ± 4.3 kg/m^2; 56.7% of participants were classified as overweight or obese (BMI ≥25 kg/m^2).

Briefly, in the ~2-hour structured laboratory visit, participants completed 11 activities selected by research staff from a larger list of options, including sedentary behaviours (e.g. lying down, writing while seating, watching television while seated), household activities/chores (e.g. dusting, making the bed, sweeping), and ambulatory/exercise activities (e.g. treadmill and overground walking, stairs, cycling) for five min each, generally in order of increasing intensity. The second visit incorporated simulated free-living/semi-structured activities which participants were free to choose in terms of order, duration, and type for 80 min, although participants were required to complete at least four activities from each category (sedentary, household/chore, ambulatory/exercise) to ensure variety in the activity types performed during the sessions. All data, including transitions and breaks (typically 1-2 minutes), were included in the present analysis. The criterion measure of time spent in MVPA was determined using direct observation of activity type (Lyden, Petruski, et al., 2014) – research assistants recorded the exact start and end time of each activity type. The 2011 Compendium of Physical Activities (Ainsworth et al., 2011) was used to determine metabolic equivalent of task (MET) for each activity, with activities requiring ≥ 3.0 METs determined to be MVPA.

Participants wore ActiGraph GT9X (Pensacola, FL; firmware version 1.1.0) and GENEActiv (Activinsights, Cambridge, UK) accelerometers on the dorsal aspect of each wrist, initialised with a sampling frequency of 60 Hz. Sampling frequency has been shown to influence the generation of activity counts. However, prior research demonstrates that use of sampling rates of 60 or 90 Hz are comparable to 30 Hz (Brønd & Arvidsson, 2016; Clevenger, Brønd, et

al., 2022). Proximal and distal positioning of the ActiGraph and GENEActiv monitors was randomised across participants but consistent between visits. All accelerometers were initialised using a common computer, which was calibrated to atomic time to ensure ease of data alignment during analysis.

Data Processing

Accelerometer data were imported into RStudio (Boston, MA; version 1.3.1056) using the 'AGread' (version 1.3.0) or the 'GENEAread' (version 2.0.9) packages. Activity counts were generated for both devices using ActiGraph's algorithm via modified code from the 'agcounts' package (version 0.1.0).

Individual classification methods for estimating time spent in MVPA were identified using recent systematic reviews (Liu et al., 2021; Migueles et al., 2017; Pfeiffer et al., 2022), the accelerometer repository (Clevenger, Montoye, et al., 2022), and literature searches (Table 1). We sought to include approaches which use a variety of inputs (raw, count data), epoch lengths (from 1- to 60-s), and model types (artificial neural networks, decision trees, two-regression, cut-points), rather than every possible available method. Of note, we did not include the Montoye et al. (2020) cut-points for non-dominant wrist vector magnitude counts because of some overlap in the data used in the present analysis. For one method (Neil-Sztramko et al., 2017), the axis was not specified, so we applied the provided cut-points to both the vertical axis and vector magnitude counts.

Three sets of models relied on the use of metrics that are orientation dependent (Montoye et al., 2016; Neil-Sztramko et al., 2017; Staudenmayer et al., 2015). The axes of the devices used in the present analyses (ActiGraph GT9X and GENEActiv) and those used in the validation studies (ActiGraph GT3X+, GENEActiv) vary in both orientation and sign direction (Supplemental Table 1). For the Staudenmayer et al. (2015) model, they indicate to use the axis

that recorded -1 *g* when the arm was hanging straight down, so we used the GT9X's x-axis, and the GENEActiv's y-axis inverted. For the Montoye et al. (2016) model, the GT9X's x- and y-axes were switched and inverted to align with the GENEActiv's y- and x-axes data, and the sign of the z-axis inverted. Finally, if the Neil-Sztramko et al. (2017) cut-points were developed for the "vertical axis," this would be equivalent to the GT9X's x-axis and the GENEActiv's y-axis when worn on the wrist.

The consensus estimate was calculated as the mean across all models developed for a wear location. Specifically, the consensus-estimate on the left wrist included 11 models, while the consensus-estimate on the right wrist included eight models. We used dominant wrist interchangeably with right wrist (and vice-versa) because our sample was almost exclusively (93%) right-hand dominant. In addition to the overall consensus-estimates, we tested three other variations of the consensus method at each wear location. First, we excluded any methods which did not include activities of daily living in their validation protocol. This was because we expect greater and more variable movement of the wrist during these types of activities as compared to locomotive or more stationary activities like watching television, which makes their inclusion important for the development of methods to classify activity intensity using wrist data. This resulted in the inclusion of seven (out of 11) methods for the left/non-dominant wrist-consensus estimate and six (out of eight) methods for the right/dominant wrist-consensus estimate. Second and third, we used the median, instead of the mean, to pool estimates for all methods or the methods which included activities of daily living in their validation protocol. Consensus estimates using the median were tested as this may be more appropriate when a few extreme estimates would affect the mean.

Statistical Analyses

All analyses were conducted in RStudio. Minutes of MVPA were compared between the criterion and the individual classification approaches and the four consensus methods using

mean absolute difference, Pearson's *r* correlation coefficient, and equivalence testing. 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. Bland-Altman plots were generated for the individual and consensus methods with the least amount of bias compared to the criterion using the 'blandr' package (version 0.6.0).

The same analytic approach was used to compare minutes of MVPA between accelerometer brands (ActiGraph versus GENEActiv) while keeping the method and wrist the same. For example, we compared the Esliger et al. (2011) cut-points applied to ActiGraph left-wrist data to the same cut-points applied to GENEActiv left-wrist data. Finally, we compared minutes of MVPA between wrists (left versus right) while keeping the device type and method the same. For example, we compared Esliger et al. (2011) cut-points applied to left-wrist data versus applied to right-wrist data. These analyses were only conducted for methods that were developed for both wrists.

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

Method	Criterion	Location	Type	Epoch (s)	Age (y)	Device	Metric	Description of MVPA determination
Dillon et al. (2016)	Indirect	NDW, DW	CP	60	18-65; NR	GA	SVM	NDW: ≥ 174.2, DW: ≥ 187.6 m*g**
Esliger et al. (2011)	Indirect	LW, RW	CP	60	40-65; NR	GA	SVM	LW: ≥ 134, RW: ≥ 92 m*g**
Hibbing et al. (2018)	Indirect	LW, RW	Two-regression	1	NR; 23.0 ± 2.3	AG GT9X	ENMO	Coefficient of variation in ENMO·1-s^{-1} determines which of two equations is used to predict METs (≥3 METs classified as MVPA). Implemented using the 'TwoRegression' package.
Hildebrand et al. (2014)	Indirect	NDW	CP	5	21-61; 34.2 ± 10.7	GA, AG GT3X+	ENMO	GENEActiv (≥93.2 m*g*), ActiGraph GT3X+ (≥100.6 m*g*)
Kwan et al. (2020)	Indirect	LW, RW	CP	60	60-73; 66.6 ± 3.5	AG GT3X+	VM counts	LW: ≥4117.1, RW: ≥4212.9 counts·min^{-1}
Lee et al. (2019)	Indirect	LW, RW	CP	60	18-26; NR	AG wGT3X-BT	VM counts	LW: ≥4514, RW: ≥4793 counts·min
Montoye et al. (2016)	DO	LW, RW	ANN	30	18-44; 22.0 ± 4.2 y	GA	Raw acceleration features	Input features include the 10, 25, 50, 75, and 90th percentiles of acceleration in each axis.
Neil-Sztramko et al. (2017)	Indirect	NDW	CP	60	22-65; 40.0 ± 14.9	AG GT3X+	Not reported	≥2199 counts·min^{-1}; women only
Rhudy et al. (2020)	Indirect	LW	CP	60	NR; 26.1 ± 9.6	AG GT9X	VM counts	≥4836 counts·min^{-1}
Sanders et al. (2019)	Indirect	NDW	CP	1	60-86; 69.6 ± 8.0	GA	ENMO	≥104 m*g*
Staudenmayer et al. (2015)	Indirect	DW	Linear regression, decision tree	15	20-39; 24.1 ± 4.5	AG GT3X+	Raw acceleration features	Standard deviation of VM of the raw acceleration and mean angle of acceleration relative to the vertical axis predicts METs (≥3 METs classified as MVPA) or activity intensity.

AG: ActiGraph; ANN: artificial neural network; CP: cut-points; DO: direct observation; DW: dominant wrist; ENMO: Euclidean norm minus one, calculated as vector magnitude of the raw acceleration minus one, with negative values rounded to zero; GA: GENEActiv; Indirect: indirect calorimetry; LW: left wrist; METs: metabolic equivalents of task; m*g*: milli-*g*; MVPA: moderate-to-vigorous physical activity; NDW: non-dominant wrist; NR: not reported; RW: right wrist; SVM: the absolute value of vector magnitude of raw acceleration minus one; VM: vector magnitude, calculated as the square root of the sum of the squared values in each axis; VA: vertical axis.*Scaled version as implemented in GGIR (*Published Cut-Points and How to Use Them in GGIR*, 2022)

Results

Comparison of Individual and Consensus Methods to Criterion

According to the criterion, participants spent 38.2 min in MVPA. On the left wrist, over-estimation was markedly worse when the Neil-Sztramko et al. (2017) cut-points were applied using vector magnitude counts (over-estimated by over 40 min), so we only retained the vertical axis analysis in the present study. Mean MVPA ranged from 40.6 to 60.4 min across the individual classification methods when using the ActiGraph, and 33.5 to 55.2 min when using the GENEActiv, while consensus estimates varied from 44.7 to 49.1 min and 41.6 to 44.9 min, respectively (Figure 1). On the right wrist, individual methods ranged from 35.1 to 73.8 min when using the ActiGraph, and 35.1 to 69.4 min when using the GENEActiv, with consensus estimates of 52.2 to 56.9 min and 51.0 to 53.9 min, respectively (Figure 2). Irrespective of accelerometer device brand, 95% of individual methods over-estimated MVPA compared to the criterion when using left-wrist data, compared to 81% on the right wrist. The bias of methods which over-estimated MVPA was greater than the bias of methods which under-estimated MVPA (Figures 1 and 2).

Mean absolute differences, correlation coefficients, and results of equivalence testing comparing individual classification methods and the consensus method with the criterion are reported in Tables 2 (ActiGraph) and 3 (GENEActiv) for the left wrist and Tables 4 (ActiGraph) and 5 (GENEActiv) for the right wrist. Across wear locations and device brands, no individual or consensus methods were statistically equivalent to the criterion.

For the left wrist, mean absolute differences between the criterion and individual methods ranged from 7.0 to 23.0 min, compared to 7.6 to 12.5 min for the consensus methods. Bias ranged from 1.4 to 22.1 min for individual methods compared to the criterion, and 3.3 to 10.8 min for the consensus methods. Correlation coefficients comparing individual methods to

the criterion ranged from 0.35 to 0.71, while comparisons of the consensus methods with the criterion resulted in correlation coefficients of 0.65 to 0.70. Bland-Altman plots (Supplemental Figure 1) show similar patterns of bias between individual and consensus methods compared to the criterion wherein there is greater over-estimation compared to the criterion in those with higher MVPA.

For the right wrist, mean absolute differences between the criterion and individual methods ranged from 8.9 to 35.8 min, compared to 13.1 to 18.6 min for the consensus methods. Bias ranged from 2.5 to 35.6 min for individual methods compared to the criterion, and 13.0 to 18.6 min for the consensus methods. Correlation coefficients comparing individual methods to the criterion ranged from 0.12 to 0.67, while comparisons of the consensus methods with the criterion resulted in correlation coefficients of 0.58 to 0.66. Full correlation matrices for both wrists are available in Supplemental Table 2. Bland-Altman plots (Supplemental Figure 2) show smaller limits of agreement for the consensus methods versus the individual methods when compared to the criterion. However, the consensus method resulted in consistent over-estimation, whereas MVPA could be over-or underestimated when using the individual methods.

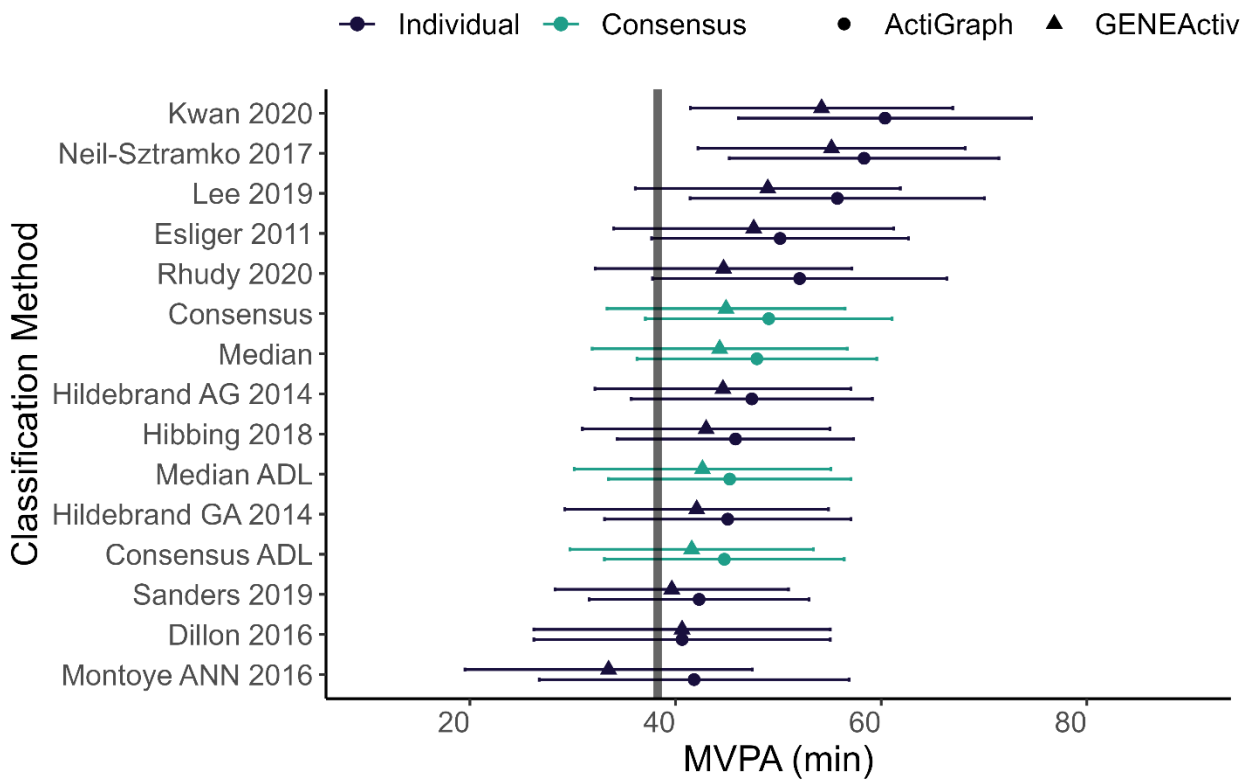
Comparison by Device Type

Mean absolute differences, correlation coefficients, and the results of equivalence testing comparing individual classification methods and the consensus method between device types are reported in Supplemental Table 3 (left) and Supplemental Table 4 (right), while an overview of the findings is reported in Table 6. On the left wrist, correlation coefficients were 0.28 to 0.99 for individual methods and 0.80 to 0.86 for consensus methods. On the right wrist, correlation coefficients were 0.24 to 0.99 for individual methods, compared to 0.70 to 0.78 for consensus methods. Across both wrists, eight (of 19) individual methods and three (of eight) consensus methods were equivalent across device types.

Comparison by Wrist

Mean absolute differences, correlation coefficients, and results of equivalence testing comparing individual classification methods and the consensus method between wrists are reported in Supplemental Table 5, while an overview of findings is reported in Table 6. Correlation coefficients were 0.58 to 0.81 for ActiGraph and 0.55 to 0.86 for GENEActiv, compared to 0.73 to 0.82 and 0.76 to 0.87, respectively, for consensus methods. No methods were statistically equivalent across wrists.

Figure 1. Comparison of minutes of moderate-to-vigorous physical activity (MVPA) according to the criterion, individual classification methods, and the consensus method using an ActiGraph and GENEActiv on the left wrist. Points (triangle or circle) represent the mean, while bars represent the standard deviation. Consensus: the mean of all 11 individual methods. Consensus ADL: the mean of seven methods which included activities of daily living in their validation protocol. Median: the median of all 11 individual methods. Median ADL: the median of seven methods which included activities of daily living in their validation protocol.



288

Figure 2. Comparison of minutes of moderate-to-vigorous physical activity (MVPA) according to the criterion, individual classification methods, and the consensus method using an ActiGraph and GENEActiv on the right wrist. Points (triangle or circle) represent the mean, while bars represent the standard deviation. Consensus: the mean of all eight individual methods. Consensus ADL: the mean of six methods which included activities of daily living in their validation protocol. Median: the median of all eight individual methods. Median ADL: the median of six methods which included activities of daily living in their validation protocol.

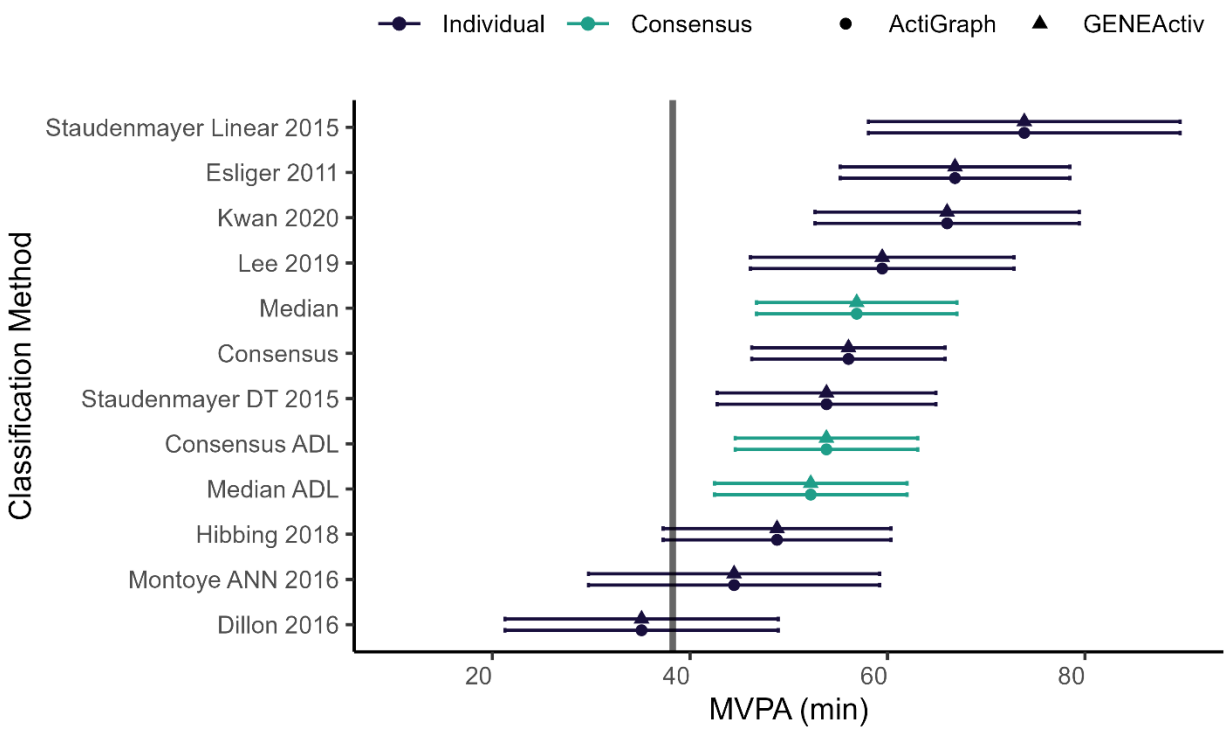


Table 2. Comparison of minutes of moderate-to-vigorous physical activity according to the criterion versus individual classification methods or the consensus method using an ActiGraph on the left wrist

	MVPA (min)				Absolute Difference				Equivalence Test	
Method	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	***r***	**Bias**	**90% CI**	**Equivalent**
Criterion	38.2	6.6	25.0	49.5	-	-	-	-	-	-
Dillon et al.(2016)	40.6	14.4	9.0	60.0	9.4	6.1	0.68	2.4	-1.0, 5.8	N
Esliger et al.(2011)	50.2	12.5	20.0	65.0	13.6	6.7	0.68	11.9	9.0, 14.8	N
Hibbing et al.(2018)	45.8	11.5	18.8	60.4	9.5	6.0	0.70	7.6	5.0, 10.2	N
Hildebrand et al. (2014) AG	45.1	12.0	16.8	60.5	10.7	6.7	0.67	9.2	6.4, 11.9	N
Hildebrand et al.(2014) GA	47.4	11.7	19.1	62.1	9.3	6.3	0.67	6.8	4.0, 9.6	N
Kwan et al.(2020)	60.4	14.2	19.0	81.0	23.0	10.0	0.58	22.1	18.5, 25.8	N
Lee et al.(2019)	55.7	14.3	18.0	76.0	18.7	9.3	0.61	17.5	13.9, 21.1	N
Montoye et al.(2016) ANN	41.8	15.1	15.0	66.0	11.2	9.1	0.35	3.6	-0.8, 7.9	N
Neil-Sztramko et al.(2017)	58.3	13.1	21.0	79.0	21.1	9.3	0.51	20.1	16.6, 23.6	N
Rhudy et al.(2020)	52.1	14.3	16.0	73.0	15.3	9.3	0.62	13.8	10.3, 17.4	N
Sanders et al.(2019)	42.3	10.7	16.5	55.4	7.1	5.4	0.66	4.0	1.5, 6.5	N
Consensus	49.1	12.0	17.3	62.7	12.5	6.5	0.66	10.8	8.0, 13.6	N
Consensus ADL	44.7	11.7	16.6	58.3	9.1	5.8	0.67	6.5	3.8, 9.2	N
Median	47.9	11.7	18.0	62.0	11.5	6.2	0.65	9.7	6.9, 12.4	N
Median ADL	45.3	11.8	16.8	60.0	9.3	6.3	0.67	7.0	4.3, 9.7	N

MVPA: moderate-to-vigorous physical activity; SD: standard deviation; AG: ActiGraph; GA: GENEActiv; ANN: artificial neural network; Consensus methods are the mean or median of multiple individual methods (Consensus: the mean of all 11 individual methods; Consensus ADL: the mean of seven methods which included activities of daily living in their validation protocol; Median: the median of all 11 individual methods; Median ADL: the median of seven methods which included activities of daily living in their validation protocol); Confidence intervals (CI) were compared to equivalence bounds of ±3.825 min to determine equivalence at $p<0.05$.

Table 3. Comparison of minutes of moderate-to-vigorous physical activity according to the criterion versus individual classification methods or the consensus method using an GENEActiv on the left wrist

	MVPA (min)				Absolute Difference				Equivalence Test	
Method	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	***r***	**Bias**	**90% CI**	**Equivalent**
Criterion	38.2	6.6	25.0	49.5	-	-	-	-	-	-
Dillon et al.(2016)	40.6	14.4	9.0	60.0	9.4	6.1	0.68	2.4	-1.0, 5.8	N
Esliger et al.(2011)	47.6	13.6	19.0	66.0	11.8	7.0	0.71	9.4	6.2, 12.5	N
Hibbing et al.(2018)	43.0	12.0	16.8	57.1	8.2	5.6	0.70	4.7	2.0, 7.5	N
Hildebrand et al. (2014) AG	42.1	12.8	15.6	56.9	9.3	6.6	0.66	6.4	3.4, 9.3	N
Hildebrand et al.(2014) GA	44.6	12.5	21.0	58.8	8.5	6.1	0.65	3.8	0.7, 6.9	N
Kwan et al.(2020)	54.2	12.8	22.0	77.0	16.3	9.0	0.68	16.0	13.0, 18.9	N
Lee et al.(2019)	49.0	12.9	17.0	74.0	12.5	7.7	0.64	10.7	7.6, 13.8	N
Montoye et al.(2016) ANN	33.5	14.0	4.0	54.5	11.4	7.8	0.37	-4.8	-8.8, -0.7	N
Neil-Sztramko et al.(2017)	55.2	13.0	23.0	74.0	17.7	9.5	0.56	16.9	13.6, 20.3	N
Rhudy et al.(2020)	44.7	12.5	14.0	70.0	9.4	6.9	0.62	6.4	3.4, 9.5	N
Sanders et al.(2019)	39.6	11.3	17.1	53.1	7.0	5.1	0.66	1.4	-1.3. 4.1	N
Consensus	44.9	11.6	16.5	59.6	9.1	5.5	0.70	6.7	4.0, 9.3	N
Consensus ADL	41.6	11.8	15.1	55.8	7.6	5.4	0.69	3.3	0.6, 6.0	N
Median	44.3	12.4	17.0	58.8	9.2	6.1	0.68	6.0	3.2, 8.9	N
Median ADL	42.6	12.5	16.8	57.1	8.4	5.9	0.68	4.4	1.5, 7.3	N

MVPA: moderate-to-vigorous physical activity; SD: standard deviation; AG: ActiGraph; GA: GENEActiv; ANN: artificial neural network; Consensus methods are the mean or median of multiple individual methods (Consensus: the mean of all 11 individual methods; Consensus ADL: the mean of seven methods which included activities of daily living in their validation protocol; Median: the median of all 11 individual methods; Median ADL: the median of seven methods which included activities of daily living in their validation protocol); Confidence intervals (CI) were compared to equivalence bounds of ±3.825 min to determine equivalence at $p<0.05$.

317 **Table 4.** Comparison of minutes of moderate-to-vigorous physical activity according to the criterion versus individual classification

318 methods or the consensus method using an ActiGraph on the right wrist

	MVPA (min)				Absolute Difference			Equivalence Test		
Method	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	***r***	**Bias**	**90% CI**	**Equivalent**
Criterion	38.2	6.6	25.0	49.5	-	-	-	-	-	-
Dillon et al.(2016)	35.1	13.8	8.0	58.0	8.9	7.2	0.61	-3.2	-6.6, 0.3	N
Esliger et al.(2011)	66.8	11.6	42.0	93.0	28.6	9.9	0.52	28.6	25.5, 31.7	N
Hibbing et al.(2018)	48.8	11.5	25.5	68.6	11.3	7.8	0.65	10.6	7.8, 13.3	N
Kwan et al.(2020)	66.0	13.4	43.0	92.0	27.8	10.7	0.61	27.8	24.5, 31.1	N
Lee et al.(2019)	59.5	13.3	36.0	87.0	21.2	10.7	0.65	21.2	17.9, 24.5	N
Montoye et al.(2016) ANN	44.5	14.7	17.5	75.5	11.8	9.7	0.32	6.2	1.9, 10.6	N
Staudenmayer et al.(2015) Linear	53.8	11.1	31.2	73.2	35.8	15.6	0.17	35.6	30.6, 40.6	N
Staudenmayer et al.(2015) DT	73.8	15.8	29.2	102.2	15.9	9.4	0.47	15.6	12.5, 18.6	N
Consensus	56.0	9.8	36.7	76.1	17.8	7.5	0.65	17.8	15.5, 20.1	N
Consensus ADL	53.8	9.2	35.7	71.6	15.6	7.3	0.62	15.6	13.3, 17.8	N
Median	56.9	10.1	36.8	79.2	18.6	7.6	0.66	18.6	16.3, 21.0	N
Median ADL	52.2	9.8	34.0	70.1	14.0	7.3	0.64	14.0	11.7, 16.3	N

319 MVPA: moderate-to-vigorous physical activity; SD: standard deviation; ANN: artificial neural network; DT: decision tree; Consensus
320 methods are the mean or median of multiple individual methods (Consensus: the mean of all eight individual methods; Consensus
321 ADL: the mean of six methods which included activities of daily living in their validation protocol; Median: the median of all eight
322 individual methods; Median ADL: the median of six methods which included activities of daily living in their validation protocol);
323 Confidence intervals (CI) were compared to equivalence bounds of ±3.825 min to determine equivalence at $p<0.05$.

324

325

326

Table 5. Comparison of minutes of moderate-to-vigorous physical activity according to the criterion versus individual classification methods or the consensus method using an GENEActiv on the right wrist

	MVPA (min)				Absolute Difference				Equivalence Test	
Method	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**	***r***	**Bias**	**90% CI**	**Equivalent**
Criterion	38.2	6.6	25.0	49.5	-	-	-	-	-	
Dillon et al.(2016)	35.1	13.8	8.0	58.0	8.9	7.2	0.61	-3.2	-6.6, 0.3	N
Esliger et al.(2011)	65.9	13.7	26.0	86.0	28.0	10.5	0.55	27.7	24.1, 31.2	N
Hibbing et al.(2018)	51.4	11.8	16.9	67.5	14.4	6.5	0.67	13.1	10.4, 15.9	N
Kwan et al.(2020)	62.4	15.0	22.0	90.0	24.8	11.3	0.55	24.1	20.2, 28.0	N
Lee et al.(2019)	55.0	14.8	21.0	84.0	18.0	10.8	0.51	16.7	12.8, 20.7	N
Montoye et al.(2016) ANN	35.8	13.0	8.0	59.5	10.1	8.5	0.24	-2.5	-6.5, 1.6	N
Staudenmayer et al.(2015) Linear	69.4	17.3	27.8	90.8	32.4	15.3	0.12	31.2	25.7, 36.7	N
Staudenmayer et al.(2015) DT	50.0	13.2	23.8	69.8	13.3	9.3	0.53	11.8	8.3, 15.2	N
Consensus	53.1	10.9	26.1	73.3	15.2	8.1	0.59	14.9	12.1, 17.6	N
Consensus ADL	51.3	10.3	27.6	68.7	13.3	7.9	0.59	13.0	10.4, 15.6	N
Median	53.9	12.1	24.0	75.8	16.3	8.7	0.58	15.7	12.6, 18.8	N
Median ADL	51.0	11.0	28.8	66.6	13.1	8.3	0.60	12.8	10.1, 15.5	N

MVPA: moderate-to-vigorous physical activity; SD: standard deviation; ANN: artificial neural network; DT: decision tree; Consensus methods are the mean or median of multiple individual methods (Consensus: the mean of all eight individual methods; Consensus ADL: the mean of six methods which included activities of daily living in their validation protocol; Median: the median of all eight individual methods; Median ADL: the median of six methods which included activities of daily living in their validation protocol); Confidence intervals (CI) were compared to equivalence bounds of ±3.825 min to determine equivalence at $p<0.05$.

Table 6. Summary of equivalence (bias in minutes of moderate-to-vigorous physical activity) for comparisons of individual and consensus methods with the criterion, across device types, and across wrists

Method	Criterion vs AG LW	Criterion vs GA LW	Criterion vs AG RW	Criterion vs GA RW	AG vs GA LW	AG vs GA RW	LW vs RW AG	LW vs RW GA	Avg
Dillon et al.(2016)	N (2.4)	N (2.4)	N (-3.2)	N (-3.2)	Y (0.0)	Y (0.0)	N (5.5)	N (5.5)	1.2
Esliger et al.(2011)	N (11.9)	N (9.4)	N (28.6)	N (27.7)	Y (2.6)	Y (0.9)	N (-16.7)	N (-18.3)	5.8
Hibbing et al.(2018)	N (7.6)	N (4.7)	N (10.6)	N (13.1)	Y (2.9)	Y (-2.6)	N (-3.0)	N (-8.4)	3.1
Hildebrand et al. (2014) AG	N (9.2)	N (6.4)	-	-	Y (2.8)	-	-	-	6.1
Hildebrand et al.(2014) GA	N (6.8)	N (3.8)	-	-	N (3.0)	-	-	-	4.5
Kwan et al.(2020)	N (22.1)	N (16.0)	N (27.8)	N (24.1)	N (6.2)	N (3.7)	N (-5.7)	N (-8.2)	10.8
Lee et al.(2019)	N (17.5)	N (10.7)	N (21.2)	N (16.7)	N (6.8)	N (4.5)	N (-3.7)	N (-6.0)	8.5
Montoye et al.(2016) ANN	N (3.6)	N (-4.8)	N (6.2)	N (-2.5)	N (8.3)	N (8.7)	N (-2.7)	N (-2.3)	1.8
Neil-Sztramko et al.(2017)	N (20.1)	N (16.9)	-	-	N (3.2)	-	-	-	13.4
Rhudy et al.(2020)	N (13.8)	N (6.4)	-	-	N (7.4)	-	-	-	9.2
Sanders et al.(2019)	N (4.0)	N (1.4)	-	-	Y (2.7)	-	-	-	2.7
Staudenmayer et al.(2015) Linear	-	-	N (35.6)	N (31.2)	-	N (4.4)	-	-	23.7
Staudenmayer et al.(2015) DT	-	-	N (15.6)	N (11.8)	-	N (3.8)	-	-	10.4
Consensus	N (10.8)	N (6.7)	N (17.8)	N (14.9)	N (4.2)	N (2.9)	N (-7.0)	N (-8.2)	5.3
Consensus ADL	N (6.5)	N (3.3)	N (15.6)	N (13.0)	N (3.2)	Y (2.6)	N (-9.1)	N (-9.7)	3.2
Median	N (9.7)	N (6.0)	N (18.6)	N (15.7)	N (3.6)	N (2.9)	N (-9.0)	N (-9.7)	4.7
Median ADL	N (7.0)	N (4.4)	N (14.0)	N (12.8)	Y (2.6)	Y (1.2)	N (-7.0)	N (-8.4)	3.3
Average	10.2	6.2	17.4	14.6	4.0	2.8	-5.8	-7.4	5.8

Dashes (-) indicate this comparison was not applicable; AVG=average; AG: ActiGraph; GA: GENEActiv; MVPA: moderate-to-vigorous physical activity; SD: standard deviation; ANN: artificial neural network; DT: decision tree; Consensus methods are the mean or median of multiple individual methods (Consensus: the mean of all individual methods; Consensus ADL: the mean of methods which included activities of daily living in their validation protocol; Median: the median of all individual methods; Median ADL: the median methods which included activities of daily living in their validation protocol)

Discussion

Use of a wrist-worn accelerometer has become increasingly popular due to improved wear compliance and easier capture of 24-h movement behaviours (Fairclough et al., 2016; Troiano et al., 2014). However, lack of agreement regarding the best way to analyse accelerometer data and inconsistencies in how accelerometer data are analysed remain fundamental barriers to surveillance of and research on how physical behaviours, such as MVPA, affect health, change over time, or vary across groups (Pedišić & Bauman, 2015). Whilst numerous analytic options exist, the present study demonstrates the difficulty in using wrist-worn accelerometer data to accurately capture time spent in MVPA in adults, as almost every existing method over-estimated MVPA compared to the criterion of direct observation.

Our first aim was to evaluate the accuracy of a consensus method, which accounts for the observation that some individual methods under- while others over-estimate MVPA, resulting in a consensus estimate that is more reflective of the criterion value. This was demonstrated in a prior study which developed a consensus method for hip-worn devices (Clevenger, Mackintosh, et al., 2022). However, in the present study, there was a systematic error in the individual methods, meaning the resultant consensus estimate was also biased. While the consensus method will never have greater error than any individual method, this highlights how this proposed method inherently captures the weaknesses of the included individual methods (“garbage in, garbage out”). The present analysis cannot identify the reason for this systematic over-estimation but it is clear that new methods with lower error are needed to characterize MVPA from wrist-based accelerometers. Given that the wrist consensus methods were not statistically equivalent to the criterion, we cannot recommend their use in future studies, in which the primary goal is to have the most accurate MVPA assessment.

However, it is pertinent to note there are still potential benefits of using a consensus method for estimating time spent in MVPA which may warrant additional research if the purpose

is to foster comparability across studies which have, or may (in the future) use, different processing methods. Specifically, the consensus method is more consistent than individual methods, as evidenced by the less variable correlations and errors across device brands and wear locations, compared to individual methods. Similarly, the range across consensus estimates is eight-times smaller than that across individual methods. Thus, it is likely that studies employing different consensus methods would enhance inter-study comparability than those employing different individual methods. Another key benefit of the consensus method is the ability to tailor, including the integration or removal of methods based on data availability, development of new methods, or updated information about the validity of earlier methods. For example, if researchers implemented a single method, such as the Dillon et al. (2016) cut-points, it would be difficult to compare findings to prior research using a different method, or to change this method if/when a better method is established. With the consensus method, even methods using completely different sets of models are comparable (Clevenger, Mackintosh, et al., 2022), and estimates are “future proofed” as methods can be added/replaced. Further, backwards comparability is afforded as individual methods could be extracted for comparison with prior studies. Finally, while individual methods are developed on relatively small, homogenous samples (e.g., all from one geographic location or age group), the consensus method may improve generalizability by pooling these methods. However, these benefits of a consensus method may be offset by the increase in analytic complexity and the associated time investment.

The data used in the present study included locomotion and simulated activities of daily living completed during both structured and semi-structured laboratory visits. As such, it could be postulated that individual methods may have been developed using only locomotive or other structured behaviours which do not involve much wrist movement, and therefore over-estimated activity intensity when applied to our data set. When we developed consensus methods which

only included studies with activities of daily living in the validation protocol, they did have an improvement of ~1-3 min in mean absolute difference compared to the criterion, but MVPA was still over-estimated. Use of the median, instead of the mean, to generate the consensus estimate did not appear to be worthwhile in the present study, with no improved accuracy in MVPA estimation. While it is expected that models will perform worse when applied to a new, independent sample (Montoye et al., 2018), further research is needed to ascertain how validation protocols can be better designed for wrist-worn accelerometer data. For example, there may be substantial variability in wrist movement when individuals perform activities of daily living, requiring larger sample sizes compared to hip-worn accelerometer validation studies.

Improving how methods are developed has been discussed frequently by researchers, with many calling for larger and more diverse samples, inclusion of a variety of activities representative of how the population spends their time, and use of both structured and unstructured (free-living) settings (Bassett et al., 2012; Keadle et al., 2019; Pfeiffer et al., 2022; Welk et al., 2005, 2019). The importance of independent sample cross-validation to better understand how models will perform in new samples and/or settings has also been highlighted (Clevenger, Montoye, et al., 2022). The present analysis also served as an independent sample cross-validation of the 19 individual methods, across two device brands. While no individual or consensus methods were statistically equivalent to the criterion, the best-performing method across device brands and wrists were the raw acceleration cut-points developed by Dillon et al. (2016) (Table 6). The Dillon et al. (2016) cut-points were developed on a convenience sample of 56 adults, 18-65 years of age, who wore GENEActiv devices and participated in sitting, standing, dish washing, floor sweeping, slow walking, fast walking and jogging for an undisclosed amount of time. Use of the Dillon et al. (2016) method is promising due to its

demonstrated validity across device brands and wrists. Further cross-validation in other independent samples, particularly in free-living, is warranted.

Another purpose of the present study was to compare between device brands and wrist wear locations. Neither wrist seemed to perform markedly better when compared to the criterion, and there were moderate correlations between individual methods that were simultaneously applied to both wrists (*r*=0.55-0.86). Researchers could consider whether this level of agreement warrants allowing participants to select which wrist they would like to wear the device on when methods are simultaneously validated for each wrist (e.g. the Dillon et al., 2016 cut-points), perhaps coupled with further research of the impact on wear compliance.

Similarly, neither device brand seemed to out-perform the other when compared to the criterion. Applying methods developed using ActiGraph counts to GENEActiv data resulted in similar comparability to the criterion as using these methods with ActiGraph data. When compared to each other, most methods were comparable between device brands worn on the same wrist, although there was marginally better agreement between ActiGraph and GENEActiv devices at the right wrist (bias 2.8 min) compared to the left wrist (4.0 min). This contradicts previous research which found that ActiGraph devices were comparable to GENEActiv at the non-dominant wrist, but that ActiGraph had a lower mean acceleration at the dominant wrist compared to the GENEActiv (Rowlands, Plekhanova, et al., 2019). It is notable that the poorest correlations were for methods which relied on the axis- and orientation- dependent raw acceleration data. Lack of clarity about how to use these methods with different device brands and/or generations limits comparability across studies. For example, the GT3X+, and GT3X-BT devices can be worn with the black cap pointed superior or inferior when in anatomical position, which influences the sign direction of the axes, yet researchers do not consistently report how the device was worn. Future research may wish to focus on the use of metrics which are axis- and orientation-independent, like vector magnitude (square root of the sum of the squared

acceleration in each axis). Moreover, manufacturers are encouraged to maintain consistency in axis direction and orientation as new generations of devices are released.

This study is not without limitations. Specifically, we had a relatively small sample with a wide age range, who only completed laboratory visits and were not observed during free-living, thereby warranting further investigation into the validity of these methods when applied to other samples or to free-living data. Individual methods may work better or worse for people of different ages; the Dillon et al. (2016) cut-points, which were the most accurate individual methods in the present analysis, used a similar age range (18-65 y) to the present study (18-79 y) which may have contributed to its accuracy. More research is also warranted that identifies whether there is a more optimal approach to weighting or selecting methods for inclusion in the consensus estimate. For example, we may use information about the demonstrated validity in an independent sample or similarity between the demographics of validation protocol with that of the sample the methods are being applied to, in order to weight the individual methods when calculating the consensus estimate. However, the consistent over-estimation of MVPA when analysing wrist accelerometer data needs to be addressed prior to further research on the optimal use of consensus methods at this wear location being conducted.

In addition to differences in sample characteristics, there may be other differences between the present study's methodology compared to that of the original validation studies which could result in the observed bias. For example, the present study used a sampling rate of 60 Hz which may not match the original validation studies; similar cross-validation studies using other sampling rates may have different findings. While individual methods were originally validated using indirect calorimetry, we elected to use a criterion of direct observation in the present study because indirect calorimetry can be difficult to employ when participants are performing various activities in succession, especially when they are unlikely to achieve a steady-state (e.g., due to the duration of the activities). The study which provided the data used

in the present analysis only required a one to two minute break between activities and activities during the semi-structured session could be quite short (two min), which would lead to a known mismatch between accelerometer-captured movement data and oxygen consumption data. Additionally, it is well known that accelerometry does not account for an individual's baseline fitness, and therefore accelerometry is not ideal for measures of relative intensity. Using direct observation reduces these differences between participants and more closely examines the association of accelerometer methods with the general intensity of an activity at the group level. Still, it is known that direct observation can underestimate time spent in MVPA compared to indirect calorimetry by ~5% (Lyden et al., 2014) – which translates to less than two minutes in the present study. Therefore, use of direct observation instead of energy expenditure as the criterion would likely not change the conclusion drawn that most wrist-worn accelerometer methods to do not accurately measure MVPA.

We elected to limit the scope of our analysis to MVPA and did not include further intensity outcomes, such as sedentary time, light physical activity, or moderate and vigorous physical activity in isolation. This was primarily due to some methods not predicting all outcomes (Rhudy et al., 2020). Similarly, it is pertinent to note that there are sedentary-specific methods because sedentary behaviour is defined by both an energy expenditure and postural component, in contrast to MVPA which is only defined by energy expenditure (Rowlands et al., 2016). Nonetheless, the increasing focus on simultaneous consideration of sedentary time, light physical activity, and MPVA warrants further research to understand whether a consensus approach is useful for estimating those outcomes.

In conclusion, better methods for estimating MVPA from wrist-worn accelerometer data are needed, given the consistent over-estimation of virtually all of the tested methods compared to a criterion of direct observation. The use of the Dillon et al. (2016) cut-points is promising, but free-living cross-validation is still needed. Whilst the wrist-worn consensus method cannot be

495 recommended at present, there are still potential benefits of this approach, such as improved
496 inter-study, -wrist, and -device brand comparability.

References

Belcher, B. R., Wolff-Hughes, D. L., Dooley, E. E., Staudenmayer, J., Berrigan, D., Eberhardt, M. S., & Troiano, R. P. (2021). US Population-referenced Percentiles for Wrist-Worn Accelerometer-derived Activity. *Medicine and Science in Sports and Exercise*.

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.

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.

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.

Clevenger, K. A., Brønd, J. C., Mackintosh, K. A., Pfeiffer, K. A., Montoye, A. H., & McNarry, M. A. (2022). Impact of ActiGraph sampling rate on free-living physical activity measurement in youth. *Physiological Measurement*, *43*(10), 105004.

Clevenger, K. A., Mackintosh, K. A., McNarry, M. A., Pfeiffer, K. A., Nelson, M. B., Bock, J. M., Imboden, M. T., Kaminsky, L. A., & Montoye, A. H. (2022). A consensus method for estimating physical activity levels in adults using accelerometry. *Journal of Sports Sciences*, 1–8.

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), 09NT01.

521 Dillon, C. B., Fitzgerald, A. P., Kearney, P. M., Perry, I. J., Rennie, K. L., Kozarski, R., & Phillips,
522 C. M. (2016). Number of days required to estimate habitual activity using wrist-worn
523 GENEActiv accelerometer: A cross-sectional study. *PloS One*, *11*(5), e0109913.
524 Ding, D., Lawson, K. D., Kolbe-Alexander, T. L., Finkelstein, E. A., Katzmarzyk, P. T., Van
525 Mechelen, W., & Pratt, M. (2016). The economic burden of physical inactivity: A global
526 analysis of major non-communicable diseases. *The Lancet*, *388*(10051), 1311–1324.
527 Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., Van
528 Hees, V. T., Trenell, M. I., & Owen, C. G. %J P. one. (2017). *Large scale population*
529 *assessment of physical activity using wrist worn accelerometers: The UK Biobank Study*.
530 *12*(2), e0169649.
531 Ellingson, L. D., Hibbing, P. R., Kim, Y., Frey-Law, L. A., Saint-Maurice, P. F., & Welk, G. J.
532 (2017). Lab-based validation of different data processing methods for wrist-worn
533 ActiGraph accelerometers in young adults. *Physiological Measurement*, *38*(6), 1045.
534 Esliger, D. W., Rowlands, A. V., Hurst, T. L., Catt, M., Murray, P., & Eston, R. G. (2011).
535 Validation of the GENEA Accelerometer. *Medicine and Science in Sports and Exercise*,
536 *43*(6), 1085–1093.
537 Fairclough, S. J., Noonan, R. J., Rowlands, A. V., Van Hees, V., Knowles, Z. R., & Boddy, L. M.
538 (2016). Wear Compliance and Activity in Children Wearing Wrist and Hip-Mounted
539 Accelerometers. *Medicine & Science in Sport & Exercise*, *48*(2), 245–253.
540 Farrahi, V., Niemelä, M., Kangas, M., Korpelainen, R., & Jämsä, T. (2019). Calibration and
541 validation of accelerometer-based activity monitors: A systematic review of machine-
542 learning approaches. *Gait & Posture*, *68*, 285–299.
543 Hibbing, P. R., Lamunion, S. R., Kaplan, A. S., & Crouter, S. E. (2018). Estimating Energy
544 Expenditure with ActiGraph GT9X Inertial Measurement Unit. *Medicine and Science in*
545 *Sports and Exercise*, *50*(5), 1093–1102.

Hildebrand, M., V. A. N. Hees VT, 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

Kraus, W. E., Powell, K. E., Haskell, W. L., Janz, K. F., Campbell, W. W., Jakicic, J. M., Troiano, R. P., Sprow, K., Torres, A., & Piercy, K. L. (2019). Physical activity, all-cause and cardiovascular mortality, and cardiovascular disease. *Medicine and Science in Sports and Exercise*, *51*(6), 1270.

Kwan, R. Y. C., Liu, J. Y. W., Lee, D., Tse, C. Y. A., & Lee, P. H. (2020). A validation study of the use of smartphones and wrist-worn ActiGraphs to measure physical activity at different levels of intensity and step rates in older people. *Gait & Posture*, *82*, 306–312.

Lee, P., & Tse, C. (2019). Calibration of wrist-worn ActiWatch 2 and ActiGraph wGT3X for assessment of physical activity in young adults. *Gait & Posture*, *68*, 141–149.

Liu, F., Wanigatunga, A. A., & Schrack, J. A. (2021). Assessment of physical activity in adults using wrist accelerometers. *Epidemiologic Reviews*, *43*(1), 65–93.

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 and Health*, *11*(4), 860–863.

McTiernan, A., Friedenreich, C. M., Katzmarzyk, P. T., Powell, K. E., Macko, R., Buchner, D., Pescatello, L. S., Bloodgood, B., Tennant, B., & Vaux-Bjerke, A. (2019). Physical activity in cancer prevention and survival: A systematic review. *Medicine and Science in Sports and Exercise*, *51*(6), 1252.

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.

Montoye, A. H., Clevenger, K. A., Pfeiffer, K. A., Nelson, M. B., Bock, J. M., Imboden, M. T., & Kaminsky, L. A. (2020). Development of cut-points for determining activity intensity from a wrist-worn ActiGraph accelerometer in free-living adults. *Journal of Sports Sciences*, *38*(22), 2569–2578.

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.

Montoye, A. H. K., Westgate, B. S., Fonley, M. R., & Pfeiffer, K. A. (2018). Cross-validation and out-of-sample testing of physical activity intensity predictions with a wrist-worn accelerometer. *Journal of Applied Physiology*, *124*(5), 1284–1293. https://doi.org/10.1152/japplphysiol.00760.2017

Montoye, Pivarnik, J. M., Mudd, L. M., Biswas, S., & Pfeiffer, K. A. (2016). Validation and comparison of accelerometers worn on the hip, thigh, and wrists for measuring physical activity and sedentary behavior. *AIMS Public Health*, *3*(2), 298.

Neil-Sztramko, S. E., Rafn, B. S., Gotay, C. C., & Campbell, K. L. (2017). Determining activity count cut-points for measurement of physical activity using the Actiwatch2 accelerometer. *Physiology & Behavior*, *173*, 95–100.

Neishabouri, A., Nguyen, J., Samuelsson, J., Guthrie, T., Biggs, M., Wyatt, J., Cross, D., Karas, M., Migueles, J. H., & Khan, S. (2022). Quantification of Acceleration as Activity Counts in ActiGraph Wearables. *Scientific Reports*, *12*(11958).

Pedišić, Ž., & Bauman, A. (2015). Accelerometer-based measures in physical activity surveillance: Current practices and issues. *British Journal of Sports Medicine*, *49*(4), 219–223.

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*.

Piercy, K. L., & Troiano, R. P. (2018). Physical activity guidelines for Americans from the US department of health and human services: Cardiovascular benefits and recommendations. *Circulation: Cardiovascular Quality and Outcomes*, *11*(11), e005263.

Published cut-points and how to use them in GGIR. (2022). https://cran.r-project.org/web/packages/GGIR/vignettes/CutPoints.html

Rhudy, M. B., Dreisbach, S. B., Moran, M. D., Ruggiero, M. J., & Veerabhadrappa, P. (2020). Cut points of the Actigraph GT9X for moderate and vigorous intensity physical activity at four different wear locations. *Journal of Sports Sciences*, *38*(5), 503–510.

Rowlands, A. V., Mirkes, E. M., Yates, T. O. M., Clemes, S., Davies, M., Khunti, K., & Edwardson, C. L. (2017). *Accelerometer-assessed physical activity in epidemiology: Are monitors equivalent?*

Rowlands, A. V., Yates, T., Olds, T. S., Davies, M., Khunti, K., & Edwardson, C. L. (2016). Wrist-Worn Accelerometer-Brand Independent Posture Classification. *Medicine and Science in Sports and Exercise*, *48*(4), 748–754. https://doi.org/10.1249/mss.0000000000000813

Saint-Maurice, P. F., Graubard, B. I., Troiano, R. P., Berrigan, D., Galuska, D. A., Fulton, J. E., & Matthews, C. E. (2022). Estimated Number of Deaths Prevented Through Increased Physical Activity Among US Adults. *JAMA Internal Medicine*, *182*(3), 349–352. https://doi.org/10.1001/jamainternmed.2021.7755

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*(11), 1270–1279.

Staudenmayer, J., He, S., Hickey, A., Sasaki, J., & Freedson, P. (2015). Methods to estimate aspects of physical activity and sedentary behavior from high-frequency wrist accelerometer measurements. *Journal of Applied Physiology*, *119*(4), 396–403.

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.

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*(13), 1019–1023.

Supplemental Table 1. Comparison of axes for each device and wear location

	GT3X+	GT3X-BT	GT9X	GENEActiv
Hip	 X: 0 *g*, Y: -1 *g*, Z: 0 *g*	 X: 0 *g*, Y: +1 *g*, Z: 0 *g*	 X: 0 g, Y: -1 *g*, Z: 0 *g*	 X: 0 *g*, Y: -1 *g*, Z: 0 *g*
RW	 X: 0 *g*, Y: -1 *g*, Z: 0 *g*  X: 0 *g*, Y: +1 *g*, Z: 0 *g*	 X: 0 *g*, Y: +1 *g*, Z: 0 *g*  X: 0 *g*, Y: -1 *g*, Z: 0 *g*	 X: -1 *g*, Y: 0 *g*, Z: 0 *g*	 X: 0 *g*, Y: +1 *g*, Z: 0 *g*
LW	 X: 0 *g*, Y: -1 *g*, Z: 0 *g*  X: 0 *g*, Y: +1 *g*, Z: 0 *g*	 X: 0 *g*, Y: +1 *g*, Z: 0 *g*  X: 0 *g*, Y: -1 *g*, Z: 0 *g*	 X: +1 *g*, Y: 0 *g*, Z: 0 *g*	 X: 0 *g*, Y: -1 *g*, Z: 0 *g*

Supplemental Table 2. Correlation matrix between criterion, consensus methods, and individual methods. Variables are generally named as the wear location (LW: left wrist or RW: right wrist), followed by the device brand (AG: ActiGraph or GA: GENEActiv), the method which may include the author's name and additional modifiers needed to specify which method when an author developed more than one method (e.g., esliger_left refers to Esliger cut-points developed for the left wrist). For consensus methods, consensus refers to the mean of all individual methods; adl is the mean of methods which included activities of daily living in their validation protocol; median is the median of all individual methods; median_adl is the median of methods which included activities of daily living in their validation protocol

**see https://osf.io/wgr6d/files/osfstorage/66eac1cd9601d2bdefb7220b

Supplemental Table 3. Comparison of minutes of moderate-to-vigorous physical activity between ActiGraph and GENEActiv devices on the left wrist

	Absolute Difference			Equivalence Test		
Method	**Mean**	**SD**	***r***	**Bias**	**90% CI**	**Equivalent**
Dillon et al.(2016)	0.0	0.0	0.99	0.0	0.0, 0.0	Y
Esliger et al.(2011)	4.0	6.0	0.87	2.6	0.5, 4.7	Y
Hibbing et al.(2018)	5.5	4.9	0.83	2.9	0.7, 5.0	Y
Hildebrand et al. (2014) AG	5.2	5.1	0.85	2.8	0.7, 4.9	Y
Hildebrand et al.(2014) GA	5.6	5.3	0.84	3.0	0.8, 5.2	N
Kwan et al.(2020)	9.2	8.7	0.66	6.2	2.7, 9.6	N
Lee et al.(2019)	10.8	8.2	0.63	6.8	3.1, 10.4	N
Montoye et al.(2016) ANN	15.4	11.5	0.28	8.3	2.9, 13.7	N
Neil-Sztramko et al.(2017)	6.3	6.3	0.80	3.2	0.6, 5.8	N
Rhudy et al.(2020)	11.5	8.5	0.58	7.4	3.6, 11.2	N
Sanders et al.(2019)	5.0	4.7	0.84	2.7	0.7, 4.6	Y
Consensus	6.6	5.3	0.80	4.2	1.8, 6.5	N
Consensus ADL	5.2	4.5	0.86	3.2	1.3, 5.1	N
Median	5.4	5.9	0.82	3.6	1.4, 5.9	N
Median ADL	4.9	5.2	0.85	2.6	0.6, 4.7	Y

MVPA: moderate-to-vigorous physical activity; SD: standard deviation; AG: ActiGraph; GA: GENEActiv; ANN: artificial neural network; Consensus methods are the mean or median of multiple individual methods (Consensus: the mean of all 11 individual methods; Consensus ADL: the mean of seven methods which included activities of daily living in their validation protocol; Median: the median of all 11 individual methods; Median ADL: the median of seven methods which included activities of daily living in their validation protocol); Confidence intervals (CI) were compared to equivalence bounds of ±5 min to determine equivalence at $p<0.05$.

Supplemental Table 4. Comparison of minutes of moderate-to-vigorous physical activity between ActiGraph and GENEActiv devices on the right wrist

	Absolute Difference			Equivalence Test		
Method	**Mean**	**SD**	***r***	**Bias**	**90% CI**	**Equivalent**
Dillon et al.(2016)	0.0	0.0	0.99	0.0	0.0, 0.0	Y
Esliger et al.(2011)	5.2	5.7	0.83	0.9	-1.5, 3.3	Y
Hibbing et al.(2018)	4.8	5.8	0.82	-2.6	-4.8, -0.4	Y
Kwan et al.(2020)	7.6	7.0	0.77	3.7	0.6, 6.7	N
Lee et al.(2019)	8.0	6.5	0.78	4.5	1.6, 7.4	N
Montoye et al.(2016) ANN	15.7	10.3	0.27	8.7	3.5, 13.9	N
Staudenmayer et al.(2015) Linear	14.0	15.4	0.24	4.4	-2.0, 10.8	N
Staudenmayer et al.(2015) DT	10.3	8.2	0.46	3.8	-0.1, 7.8	N
Consensus	6.0	4.5	0.78	2.9	0.8, 5.1	N
Consensus ADL	5.8	4.1	0.77	2.6	0.5, 4.6	Y
Median	6.8	5.6	0.73	2.9	0.4, 5.5	N
Median ADL	5.9	5.6	0.70	1.2	-1.3, 3.7	Y

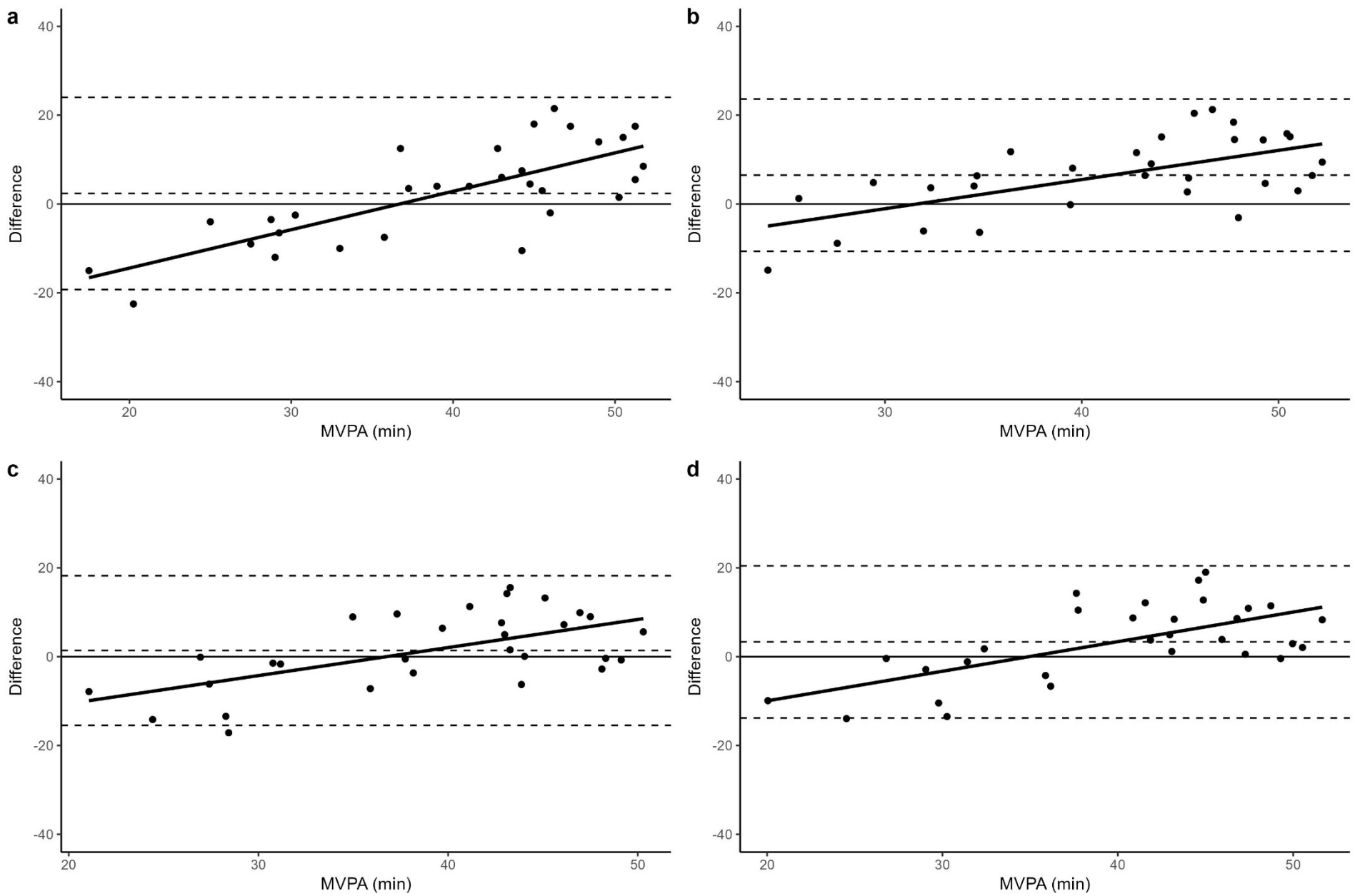
MVPA: moderate-to-vigorous physical activity; SD: standard deviation; ANN: artificial neural network; DT: decision tree; Consensus methods are the mean or median of multiple individual methods (Consensus: the mean of all eight individual methods; Consensus ADL: the mean of six methods which included activities of daily living in their validation protocol; Median: the median of all eight individual methods; Median ADL: the median of six methods which included activities of daily living in their validation protocol); Confidence intervals (CI) were compared to equivalence bounds of ±5 min to determine equivalence at $p<0.05$.

Supplemental Table 5. Comparison of minutes of moderate-to-vigorous physical activity between left and right wrists

	ActiGraph						GENEActiv					
	Absolute Difference			Equivalence Test			Absolute Difference			Equivalence Test		
Method	**Mean**	**SD**	***r***	**Bias**	**90% CI**	**Equivalent**	**Mean**	**SD**	***r***	**Bias**	**90% CI**	**Equivalent**
Dillon et al.(2016)	8.0	6.7	0.80	5.5	2.8, 8.3	N	8.0	6.7	0.80	5.5	2.8, 8.3	N
Esliger et al.(2011)	16.7	9.8	0.67	-16.7	-19.7, -13.6	N	18.3	11.6	0.64	-18.3	-21.9, -14.7	N
Hibbing et al.(2018)	5.9	4.8	0.81	-3.0	-5.2, -0.8	N	8.7	5.8	0.86	-8.4	-10.3, -6.5	N
Kwan et al.(2020)	10.9	8.3	0.58	-5.7	-9.6, -1.7	N	11.5	9.4	0.60	-8.2	-12.0, -4.3	N
Lee et al.(2019)	9.9	7.2	0.64	-3.7	-7.4, -0.1	N	10.9	8.8	0.58	-6.0	-10.0, -2.0	N
Montoye et al.(2016) ANN	7.6	6.2	0.79	-2.7	-5.6, 0.3	N	10.8	7.1	0.55	-2.3	-6.3, 1.7	N
Consensus	7.7	6.7	0.79	-7.0	-9.3, -4.7	N	8.5	7.2	0.77	-8.2	-10.6, -5.8	N
Consensus ADL	9.2	6.6	0.82	-9.1	-11.2, -7.0	N	9.7	6.6	0.83	-9.7	-11.8, -7.6	N
Median	9.5	7.5	0.73	-9.0	-11.5, -6.4	N	10.0	8.0	0.76	-9.7	-12.3, -7.0	N
Median ADL	7.7	6.5	0.78	-7.0	-9.3, -4.7	N	8.6	6.0	0.87	-8.4	-10.4, -6.5	N

MVPA: moderate-to-vigorous physical activity; SD: standard deviation; ANN: artificial neural network; DT: decision tree; Consensus methods are the mean or median of multiple individual methods (Consensus: the mean of all individual methods; Consensus ADL: the mean of methods which included activities of daily living in their validation protocol; Median: the median of all individual methods; Median ADL: the median methods which included activities of daily living in their validation protocol); Confidence intervals (CI) were compared to equivalence bounds of ±5 min to determine equivalence at $p<0.05$.

Supplemental Figure 1. Bland-Altman plots showing the difference in minutes of moderate-to-vigorous physical activity (MVPA) according to the critterion of direct observation compared to a) Dillon cut-points applied to an ActiGraph accelerometer, b) consensus activities of daily living method applied to an ActiGraph accelerometer, c) Sanders cut-points applied to a GENEActiv accelerometer, and d) consensus activities of daily living method applied to a GENEActiv accelerometer, all at the left wrist.


a
b
c
d
Difference
MVPA (min)
40
20
0
-20
-40
20
30
40
50

Supplemental Figure 2. Bland-Altman plots showing the difference in minutes of moderate-to-vigorous physical activity (MVPA) according to the critterion of direct observation compared to a) Dillon cut-points applied to an ActiGraph accelerometer, b) median consensus activities of daily living method applied to an ActiGraph accelerometer, c) Montoye artificial neural network applied to a GENEActiv accelerometer, and d) median

consensus activities of daily living0020method applied to a GENEActiv accelerometer, all at the right wrist.