PLOS ONE

RESEARCH ARTICLE

Comparison of raw accelerometry data from ActiGraph, Apple Watch, Garmin, and Fitbit using a mechanical shaker table

James W. White, III[1]*, Olivia L. Finnegan[1], Nick Tindall[1], Srihari Nelakuditi[2], David E. Brown, III[3], Russell R. Pate[1], Gregory J. Welk[4], Massimiliano de Zambotti[5], Rahul Ghosal[6], Yuan Wang[6], Sarah Burkart[1], Elizabeth L. Adams[1], Mvs Chandrashekhar[2], Bridget Armstrong[1], Michael W. Beets[1], R. Glenn Weaver[1]

1 Department of Exercise Science, University of South Carolina, Columbia, SC, United States of America, **2** Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, United States of America, **3** Division of Pediatric Pulmonology, Pediatric Sleep Medicine, Prisma Health Richland Hospital, Columbia, SC, United States of America, **4** Department of Kinesiology, Iowa State University, Ames, IA, United States of America, **5** SRI International, Menlo Park, CA, United States of America, **6** Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, United States of America

* jww4@email.sc.edu

OPEN ACCESS

Citation: White JW, III, Finnegan OL, Tindall N, Nelakuditi S, Brown DE, III, Pate RR, et al. (2024) Comparison of raw accelerometry data from ActiGraph, Apple Watch, Garmin, and Fitbit using a mechanical shaker table. PLoS ONE 19(3): e0286898. https://doi.org/10.1371/journal.pone.0286898

Editor: Yosuke Yamada, National Institute of Biomedical Innovation Health and Nutrition, JAPAN

Received: May 23, 2023

Accepted: February 12, 2024

Published: March 29, 2024

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Research reported in this publication was supported in part by the National Institute of Diabetes and Digestive and Kidney Diseases Award Number R01DK129215 (RGW). JWW was supported by the National Institute of Diabetes and Digestive and Kidney Diseases Award Number F31DK136205. OLF was supported by the National

Abstract

The purpose of this study was to evaluate the reliability and validity of the raw accelerometry output from research-grade and consumer wearable devices compared to accelerations produced by a mechanical shaker table. Raw accelerometry data from a total of 40 devices (i.e., n = 10 ActiGraph wGT3X-BT, n = 10 Apple Watch Series 7, n = 10 Garmin Vivoactive 4S, and n = 10 Fitbit Sense) were compared to reference accelerations produced by an orbital shaker table at speeds ranging from 0.6 Hz (4.4 milligravity-mg) to 3.2 Hz (124.7mg). Two-way random effects absolute intraclass correlation coefficients (ICC) tested inter-device reliability. Pearson product moment, Lin's concordance correlation coefficient (CCC), absolute error, mean bias, and equivalence testing were calculated to assess the validity between the raw estimates from the devices and the reference metric. Estimates from Apple, ActiGraph, Garmin, and Fitbit were reliable, with ICCs = 0.99, 0.97, 0.88, and 0.88, respectively. Estimates from ActiGraph, Apple, and Fitbit devices exhibited excellent concordance with the reference CCCs = 0.88, 0.83, and 0.85, respectively, while estimates from Garmin exhibited moderate concordance CCC = 0.59 based on the mean aggregation method. ActiGraph, Apple, and Fitbit produced similar absolute errors = 16.9mg, 21.6mg, and 22.0mg, respectively, while Garmin produced higher absolute error = 32.5mg compared to the reference. ActiGraph produced the lowest mean bias 0.0mg (95%CI = -40.0, 41.0). Equivalence testing revealed raw accelerometry data from all devices were not statistically significantly within the equivalence bounds of the shaker speed. Findings from this study provide evidence that raw accelerometry data from Apple, Garmin, and Fitbit devices can be used to reliably estimate movement; however, no estimates were statistically significantly equivalent to the reference. Future studies could explore device-agnostic and harmonization

Institute of General Medical Sciences Award Number T32GM081740 and T32GM145226. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. https://www.niddk.nih.gov/research-funding.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: Unrelated to this work Dr. Weaver and Dr. Armstrong report board membership and ownership shares in Trackster LLC. Unrelated to this work Dr. de Zambotti reports grants from Noctrix Health and Verily Life Science LLC (Alphabet Inc.), and is a co-founder and Chief Scientific Officer at Lisa Health Inc. and has ownership of shares in Lisa Health.

methods for estimating physical activity using the raw accelerometry signals from the consumer wearables studied herein.

Introduction

Over the past 20 years, device-based assessment of physical activity has improved due to the introduction of wearable monitors, such as accelerometers. Wearable monitors provide device-based estimates of movement and overcome recall and desirability bias that may hamper self-reported measures of physical activity [1, 2]. Best practice recommendations for using accelerometers have shifted over the last decade from traditional activity counts (accelerations per a given epoch) [3] to using raw accelerometry data from accelerometers (i.e., x-, y-, and z-axis accelerometry data in g's typically collected multiple times per second) to estimate physical activity [4].

Consumer wearables (e.g., Apple Watch, Fitbit, Garmin) are increasingly popular measurement tools for assessing physical activity. Not only are these devices equipped with accelerometers to capture movement, but they are also unobtrusive and designed to be worn on the wrist, targeted for comfort and style, affordable for consumers, rechargeable, waterproof, and can be designed for children [5–8]. Technological advances allow consumer wearables to also frequently have extended battery life (i.e., up to 54 days) [9] and remote data capture and monitoring. For these reasons, there has been a multitude of measurement studies that have explored the validity of physical activity estimates produced by consumer wearables [10, 11].

However, these studies are limited because they rely on estimates of physical activity that are derived from proprietary algorithms developed by the companies that produce these devices (e.g., Apple, Garmin, Fitbit, etc.). This is a key limitation because these algorithms are unavailable for review by researchers [12]. The drawbacks of estimating physical activity based on proprietary algorithms are that it is unclear whether best practice recommendations were used to develop these algorithms, and the algorithms could be updated by these companies at any time unbeknownst to the user [13, 14]. Thus, estimates of physical activity collected from the same device across time may provide different estimates of activity due solely to changes in the underlying algorithms that produce these metrics [13, 14].

An alternative, device-agnostic or monitor-independent approach may address these limitations by enabling data from any device to be processed using the same algorithm or processing methodology [15, 16]. A device-agnostic approach is a realistic possibility as consumer wearables have released application programming interfaces (API) that allow access to the raw accelerometry data (i.e., x, y, z axis readings collected by these devices) [17]. This has the potential to increase the comparability of physical activity estimates across time and between different consumer wearables and research-grade devices.

A necessary first step to applying a device-agnostic approach to raw accelerometry data collected by consumer wearables is to conduct mechanical signal testing of the data via controlled protocols [18]. This testing allows for the evaluation of device signals and their response to known stimuli. It also allows for the evaluation of reliability and validity of the raw acceleration output from consumer wearables without the influence of human variation [18]. It is also useful to evaluate the raw acceleration output from research-grade devices herein because it allows us to compare the acceleration output from research-grade and consumer wearables on the same metric, when compared to more direct estimates of acceleration from a mechanical shaker table. Therefore, this study will evaluate the between-device reliability and validity of

the raw acceleration output from research-grade and consumer wearable devices, compared to accelerations produced by a mechanical shaker table as the reference metric. While studies have previously examined research-grade accelerometers with this methodology [19, 20], this is among the first studies to report shaker table outcomes evaluating the raw accelerometry data from consumer-grade devices.

Methods

Raw accelerometry data from a total of 40 devices were evaluated in this study. The research-grade devices included n = 10 ActiGraph wGT3X-BT (ActiGraph; ActiGraph LLC Pensacola, FL). The consumer wearable devices included n = 10 Apple Watch Series 7 (Apple; Apple Technology Company, Cupertino, CA), n = 10 Garmin Viovactive 4S (Garmin; Garmin Ltd., Olathe, KS), and n = 10 Fitbit Sense (Fitbit; Google LLC, San Francisco, CA). Inter-device reliability and validity of raw accelerations for all devices were tested, with accelerations produced by a mechanical shaker table (Scientific Industries, Bohemia, NY; Mini-300 Orbital-Genie, Model 1500) as the reference. Each device was securely mounted directly to the twin ratcheting clamps of a mechanical shaker table (S1 Fig) that produces controlled oscillations at frequencies between approximately f_{shaker} = 0.6 and 5 Hertz (Hz). We converted f_{shaker} in Hz to acceleration using the expression for centripetal acceleration, $a_{orbital} = v^2/r_{orbital}$ [21], where $r_{orbital}$ is the radius of rotation for the orbital shaker $r_{orbital}$. From the manual for this particular shaker (supplementary https://cdn.shopify.com/s/files/1/0489/6990/8374/files/SI-M1600_Manual.pdf?v=1617998279), the specified diameter of the orbit is $2r_{orbital}$ = 1.9cm and the rotational speed is given by $v = 2\pi r_{orbital} f_{shaker}$, since for each complete cycle of 2π radians, the table traverses a distance of circumference $2\pi r_{orbital}$ in time $1/f_{shaker}$. In other words:

$$a_{orbital}(cm/s^2) = 4\pi^2 r_{orbital} f_{shaker}^2$$

to convert this acceleration to units of earth's gravity (g's), divide $a_{orbital}$ by 9.81cm/s^2.

A total of five devices were placed on the shaker table at once. Serial number/device ID and position of devices (numbered 1 to 5 from left to right) were recorded for all devices. Prior to each trial, the shaker table was placed on a level surface (i.e., floor); time from each device was recorded at the second level.

Device software

ActiGraphs were initialized to provide output from each directional axis using ActiLife software (version 6.13.4; ActiGraph LLC, Pensacola, FL). Garmin devices were initialized, and data were recorded in RawLogger (version 1.0.20211201a) and exported through Garmin Connect software™. Apple devices were initialized, and data were recorded in SensorLog (version 5.2) and exported into comma-separated values (CSV) files via Health Auto Export (version 6.3). RawLogger and SensorLog are user-written apps that leverage the device-specific Application Programming Interface (API) to collect the underlying sensor data on the respective devices. RawLogger is available for download through the Connect IQ™ store on the Garmin Connect™ app, and SensorLog and Health Auto Export are available for download through the App Store. The research team developed a custom Fitbit app (Slog) leveraging the Fitbit API for the same purpose, and Fitbit devices were initialized, and data were recorded and exported through this app. The GitHub code for the custom Fitbit app is available athttps://github.com/ACOI-UofSC/Slog_HR. Sampling frequencies from 25 Hz to 100 Hz were recorded based on the capabilities of the ActiGraph (100 Hz), Apple (100 Hz), Garmin (25 Hz), and Fitbit (50 Hz).

Reliability testing

Reliability testing included five identical devices mounted side-by-side (e.g., 5 ActiGraph devices) positioned 1–5 from left to right. Each device was tested for a total of 10 trials (5 trials at 0.6 Hz and 3.2 Hz) that lasted 2 minutes each [20]. A 15-second rest period took place at the beginning and end of each trial. Thus, it took ten minutes and 30 seconds to test 5 devices at one speed. The time of the 15-second rest periods and the trial start and end time were recorded based on device time. A minimum of 20 trials were conducted for each device brand, totaling 80 trials. Trials with missing data due to device malfunction: Apple (n = 20) and Fitbit (n = 10) were repeated to ensure that raw acceleration data from all devices could be analyzed. No trials had to be repeated for ActiGraph and Garmin devices.

Validity testing

For validity testing, five identical devices were mounted side-by-side until all devices were run through the validity trials. The trials lasted 14 minutes and 30 seconds. Consistent with past validation studies [20, 22], each trial began with a 15-second rest period (i.e., no movement) followed by a standardized series of oscillations at seven frequencies (i.e., 3.2 Hz, 2.8 Hz, 2.4 Hz, 1.9 Hz, 1.5 Hz, 1.0 Hz, 0.6 Hz) lasting two minutes each. These frequencies were chosen because they are consistent with human movement ranging from 1.5 to 16 mph [23]. The start and stop times were noted at each frequency for both research-grade and consumer wearable devices. Each trial ended with another 15-second rest period. A minimum of 2 trials were conducted for each device brand, totaling 8 trials. Trials/devices with missing data due to device malfunction: Apple (n = 4) and Fitbit (n = 1) or shaker table malfunction (n = 1) were repeated to minimize missing data; no trials had to be repeated for ActiGraph or Garmin devices. Following all testing, raw acceleration data for both research-grade and consumer wearable devices were downloaded and converted to a CSV file using ActiLife software and the device-specific user-written apps, respectively.

Sample size considerations

A sample size of 10 was selected to be consistent with previous research [19] and to provide reasonable variability within and between devices. Further, by selecting 10 devices the study was adequately powered to detect equivalence bounds of ±10% from the shaker table speed. Power is determined for an equivalence test by identifying the likelihood that the difference between two estimates is within prespecified equivalence bounds [24]. Power is then determined based upon the smallest acceptable width of the equivalence bounds. Power was calculated to detect equivalence between devices for estimates of light activity and MVPA. With a sample of 10 of each accelerometer, assuming an alpha of 0.05, and a standard deviation of the difference 10%, the study was adequately powered (power = 0.8) to detect equivalence bounds from -10% to 10% difference using standard statistical tests.

Data processing

Raw acceleration data from all devices (i.e., ActiGraph, Apple, Garmin, and Fitbit) were extracted from the middle minute of each 2-minute oscillation frequency. Consistent with past research, Euclidean Norm Minus One (ENMO) was calculated [25–28]. All values were multiplied by 1000 (milligravity-mg) to be consistent with published intensity thresholds based on the GGIR package for accelerometry in R statistical software [29]. Data were aggregated to the second level by extracting the mean and root mean square (RMS) value for each second for all devices for ENMO. We calculated both mean and RMS, as both have been calculated

previously, suggesting that there is no consensus on aggregation methods for raw accelerometry data [20, 22, 30].

Correlation coefficients

Two-way random effects absolute intraclass correlation coefficients (ICC) were calculated to assess reliability for all devices. ICC values less than 0.50 were defined as poor reliability, between 0.50 and 0.75 as moderate reliability, between 0.75 and 0.90 as good reliability, and greater than 0.90 as excellent reliability [31]. Prior to statistical analyses for validity testing, descriptive means and standard deviations for the mean and RMS were calculated across devices for each speed ranging from 0.6 to 3.2 Hz. For the validity testing, Pearson product moment *(r)* and Lin's concordance correlation coefficient (CCC) were calculated to assess correlation and agreement of raw acceleration data from ActiGraph and consumer wearable devices compared to the reference (i.e., acceleration from the shaker table) [32]. Pearson product moment interpretations were defined based on Dancey and Reidy [33], and Lin's concordance correlation coefficient was defined similarly based on recommendations from Altman (1991), with coefficients less than 0.20 as poor and greater than 0.80 as excellent [34].

Discrepancy analyses

An absolute error was calculated to assess the magnitude of the error between the reference metric and the raw acceleration data from ActiGraph and consumer wearable devices. The mean bias was also calculated to assess whether the raw acceleration output from ActiGraph and consumer wearable devices over- or underestimated acceleration output compared to the reference metric. Raw acceleration data from one ActiGraph (ID = 210) was eliminated because the device was faulty and provided implausible acceleration values (all ENMO values were below 0). Thus, there were (N = 3,780) observations for ActiGraph, whereas Apple and Garmin devices contributed (N = 4,200) observations. Missing data were present across all Fitbit devices except two, which contributed to (N = 3,975) observations for Fitbit.

Equivalence testing

Following the discrepancy analyses above, the Two-One-Sided-Tests method [35] was adopted to assess the equivalence of the raw accelerometry data collected from the accelerometers compared with accelerations from the shaker table [36]. For equivalence testing, the null hypothesis is that the raw data collected via the accelerometers and the shaker table speeds are not equivalent. To test this 90% equivalence bounds are required [37]. An equivalence zone of ±10% was adopted based upon previous work and industry standards [37, 38]. Thus, should the 90% confidence interval of the accelerometer data fall completely within ±10% of the shaker table speed, equivalence is concluded. The 'tost' command in Stata was used to complete all equivalence analyses.

Results

For reliability, ICCs (95% confidence intervals) are presented for the raw acceleration data from all devices for both aggregation methods (i.e., mean and RMS) for all devices in Table 1. The ICCs for ActiGraph were 0.97 (0.92, 0.99) and 0.97 (0.93, 0.98) for the mean and RMS aggregation methods, respectively. The ICCs for Apple were 0.99 (0.99, 0.99) and 0.99 (0.99, 1.00) for the mean and RMS, respectively. The ICCs for Garmin were 0.88 (0.82, 0.92) and 0.90 (0.85, 0.93) for the mean and RMS aggregation methods, respectively. The ICCs for Fitbit

Table 1. Summary of intraclass correlation coefficients for all devices aggregated based on the mean and root mean square.

Device	Mean	95CI	RMS	95CI
ActiGraph	0.97	(0.92, 0.99)	0.97	(0.93, 0.98)
Apple	0.99	(0.99, 0.99)	0.99	(0.99, 1.00)
Garmin	0.88	(0.82, 0.92)	0.90	(0.85, 0.93)
Fitbit	0.88	(0.86, 0.89)	0.87	(0.85, 0.88)

Abbreviations: "95CI" 95% confidence interval, "RMS" root mean square

https://doi.org/10.1371/journal.pone.0286898.t001

were 0.88 (0.86, 0.89) and 0.87 (0.85, 0.88) for the mean and RMS aggregation methods, respectively.

For validity, a summary table of outcomes based on the raw acceleration data from all devices is presented in Table 2. Fig 1 shows the raw signals with baselines for all four monitors at 1.9 Hz. Fig 2 shows the concordance of the raw acceleration data from all devices compared to the reference metric. Fig 3 shows the absolute error of the raw acceleration data from all devices compared to the reference metric. Fig 4 are Bland-Altman plots based on the estimated mean ENMO for each device compared to accelerations from the reference metric. Fig 5 are Bland-Altman plots based on the estimated RMS ENMO for each device compared to accelerations from the reference metric.

Pearson product moment correlations between raw accelerometry estimates for ActiGraph and the reference metric were $r = 0.88$ and $r = 0.89$ for the mean and RMS aggregation methods, respectively. CCCs (95% confidence intervals) when compared to the shaker table were $r_c = 0.88$ (0.87, 0.88) and $r_c = 0.88$ (0.88, 0.89) for the mean and RMS aggregation methods, respectively. Mean bias (95% confidence intervals) was 0.0mg (-40.0, 41.0) and 4.0mg (-36.0, 44.0), and absolute error was 16.9mg and 16.7mg for the mean and RMS aggregation methods, respectively.

Pearson product moment correlations between raw accelerometry estimates for Apple and the reference metric were $r = 0.94$ and $r = 0.94$ for the mean and RMS aggregation methods, respectively. CCCs when compared to the shaker table were $r_c = 0.83$ (0.82, 0.83) and $r_c = 0.90$ (0.89, 0.90) for the mean and RMS aggregation methods, respectively. Mean bias (95% confidence intervals) was -21.0mg (-50.0, 7.0) and -12.0mg (-45.0, 21.0), and absolute error was 21.6mg and 18.0mg for the mean and RMS aggregation methods, respectively.

Pearson product moment correlations between raw accelerometry estimates for Garmin and the reference metric were $r = 0.79$ and $r = 0.84$ for the mean and RMS aggregation

Table 2. Summary statistics for all devices based on the mean and root mean square aggregation methods.

	Devices	ActiGraph	Apple	Garmin	Fitbit
Mean	Observations	3,780	4,200	4,200	3,975
	Mean (mg)	54.4	32.7	23.8	46.1
	SD (mg)	41.5	41.0	34.1	57.4
	Pearson's r	0.88	0.94	0.79	0.91
Root Mean Square	Observations	3,780	4,200	4,200	3,975
	Mean (mg)	58.1	41.8	29.0	58.8
	SD (mg)	45.0	48.9	37.9	71.8
	Pearson's r	0.89	0.94	0.84	0.92

Abbreviations: "SD" standard deviation, "mg" = milligravity

https://doi.org/10.1371/journal.pone.0286898.t002

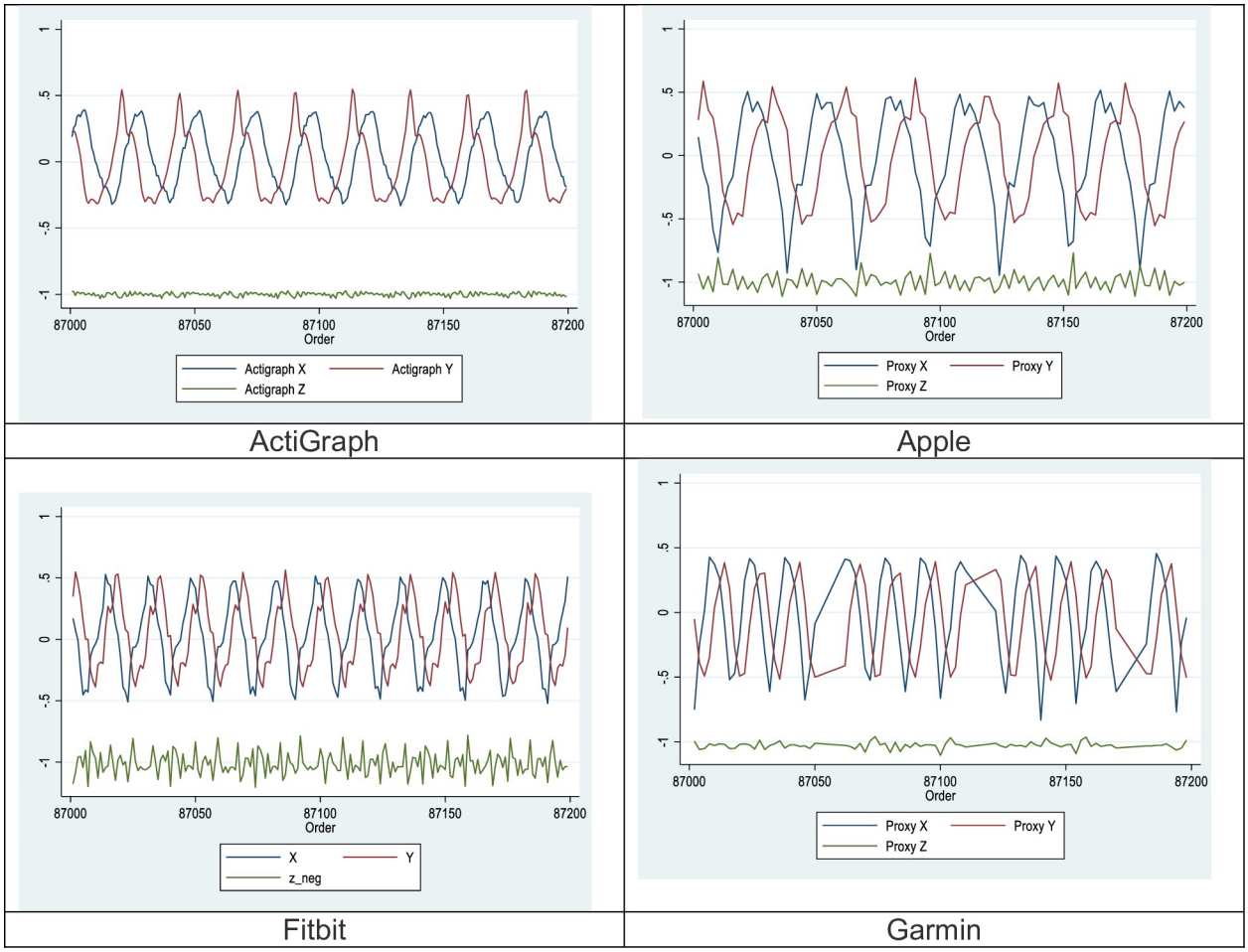

Fig 1. Raw signals with baselines for all four monitors at 1.9 Hz.

https://doi.org/10.1371/journal.pone.0286898.g001

methods, respectively. CCCs when compared to the shaker table were $r_c = 0.59$ (0.58, 0.60) and $r_c = 0.70$ (0.69, 0.71) for the mean and RMS aggregation methods, respectively. Mean bias (95% confidence intervals) was -30.0mg (-80.0, 19.0) and -25.0mg (-69.0, 19.0), and absolute error was 32.5mg and 28.1mg for the mean and RMS aggregation methods, respectively.

Pearson product moment correlations between raw accelerometry estimates for Fitbit and the reference metric were $r = 0.91$ and $r = 0.92$ for the mean and RMS aggregation methods, respectively. CCCs when compared to the shaker table were $r_c = 0.85$ (0.84, 0.86) and $r_c = 0.79$ (0.78, 0.80) for the mean and RMS aggregation methods, respectively. Mean bias (95% confidence intervals) was -8.0mg (-59.0, 44.0) and 5.0mg (-69.0, 79.0), and absolute error was 22.0mg and 24.2mg for the mean and RMS aggregation methods, respectively.

Findings from the equivalence tests between the raw acceleration estimates from all devices and the reference metric are presented in Table 3. No device estimates were found to be statistically significantly equivalent no matter the aggregation method when compared to the reference metric. For ActiGraph, mean differences were -12.9 and -9.1 based on the mean and RMS aggregation methods, respectively. For Apple, mean differences were -29.6 and -20.5 based on the mean and RMS aggregation methods, respectively. For Garmin, mean differences were -38.5 and -34.2 based on the mean and RMS aggregation methods, respectively. For

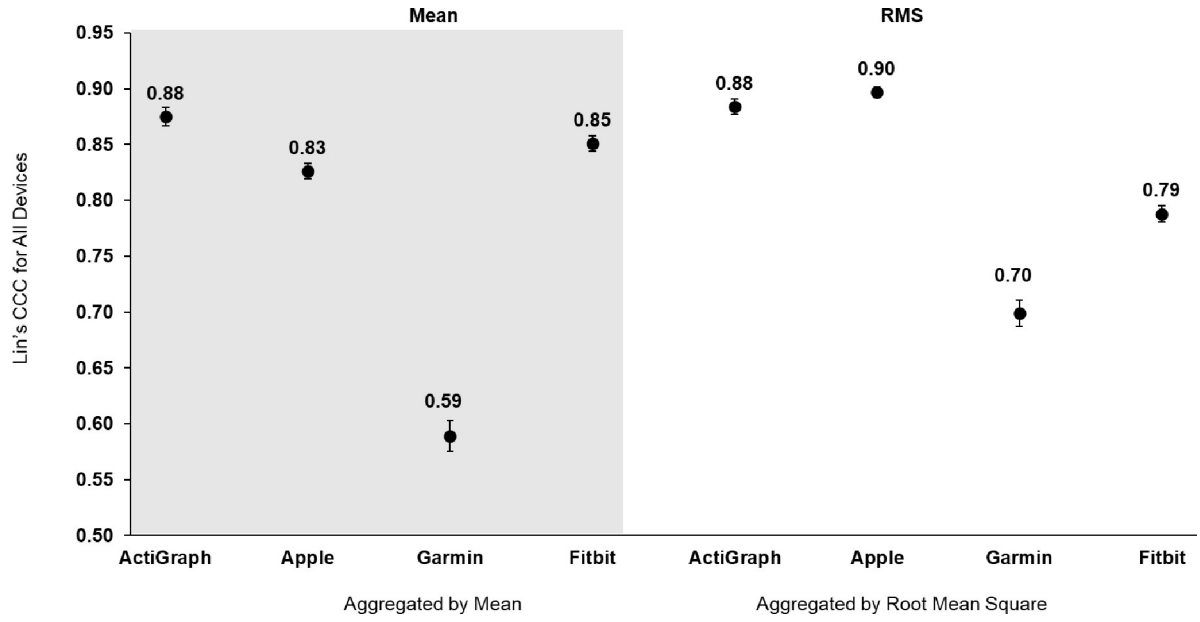

Fig 2. Lin's concordance correlation coefficient of the raw acceleration data from all devices compared to the accelerations produced by a mechanical shaker table. Error bars represent standard error.

https://doi.org/10.1371/journal.pone.0286898.g002

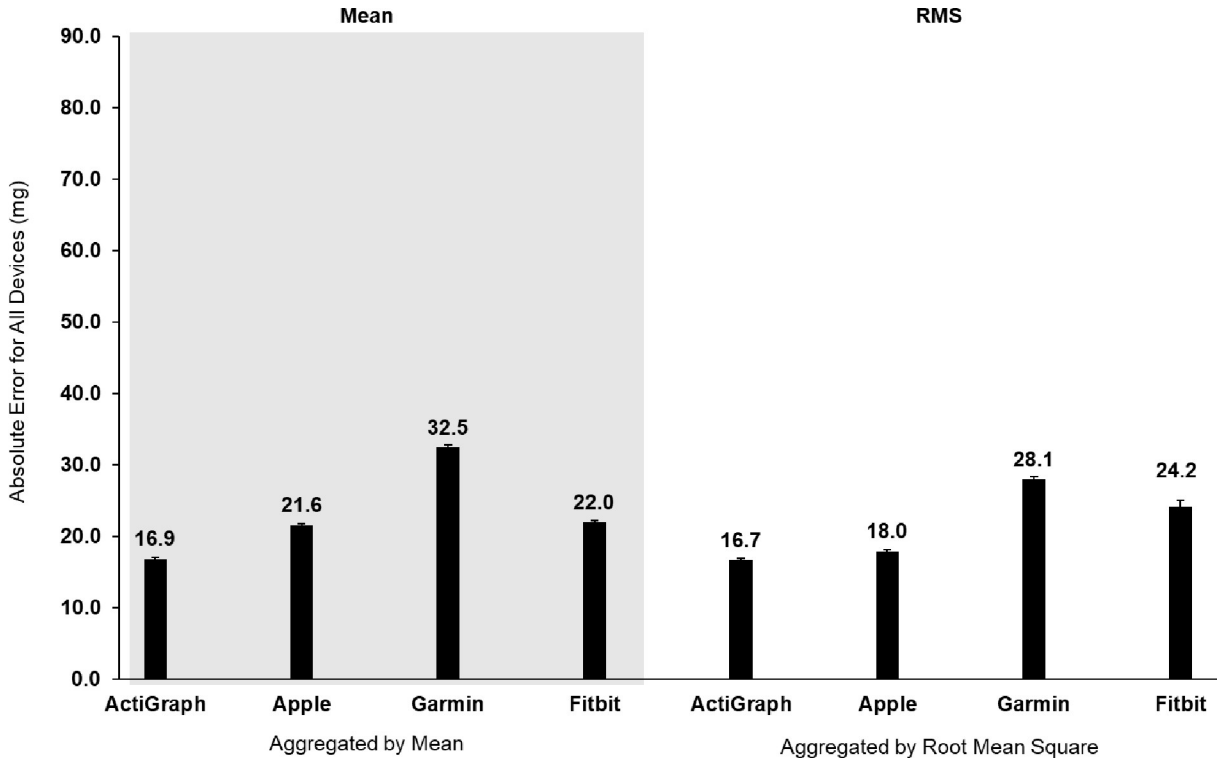

Fig 3. Absolute error of the raw acceleration data from all devices compared to the accelerations produced by a mechanical shaker table. Error bars represent standard error.

https://doi.org/10.1371/journal.pone.0286898.g003

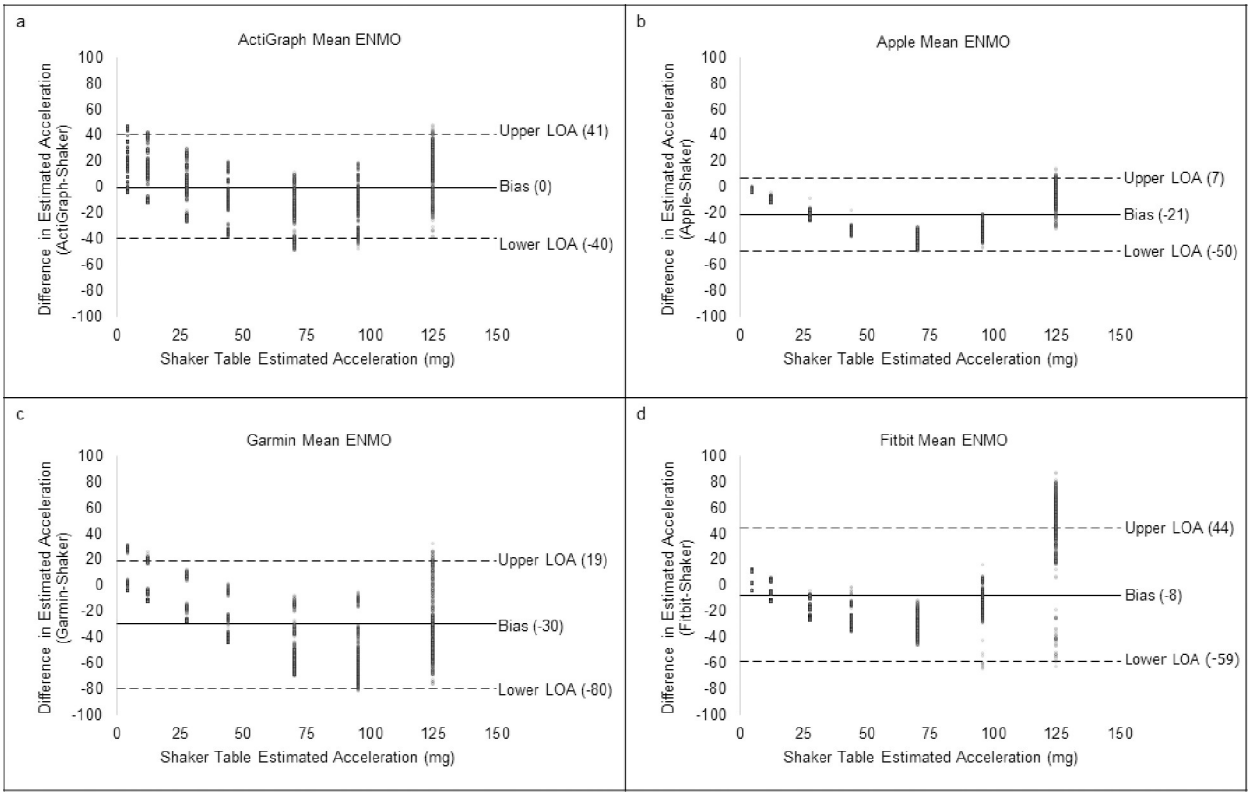

Fig 4. Bland-Altman plots of estimated mean ENMO from all devices compared to estimated shaker table acceleration.

https://doi.org/10.1371/journal.pone.0286898.g004

Fitbit, the mean differences were -16.2 and -3.5 based on the mean and RMS aggregation methods, respectively.

Discussion

The aim of this study was to evaluate the between-device reliability and validity of the raw acceleration output from research-grade (i.e., ActiGraph wGT3X-BT) and consumer wearable devices (i.e., Apple Watch Series 7, Garmin Vivoactive 4S, and Fitbit Sense) compared to accelerations produced by a mechanical shaker table. The raw acceleration data collected from all devices exhibited good-to-excellent between-device reliability based on the mean and RMS aggregation methods. For validity, the raw acceleration data from all devices exhibited a strong positive correlation to the reference metric with moderate-to-excellent concordance no matter the aggregation method. Except for Garmin, the raw acceleration data collected from consumer wearables demonstrated absolute errors with the reference metric that were similar to ActiGraph. However, equivalence testing revealed raw accelerometry data from all devices were not significantly within the equivalence bounds of the shaker speed. Moreover, the raw acceleration data collected from consumer wearables underestimated acceleration output to a greater degree than ActiGraph, when compared to the accelerations produced by the mechanical shaker table. Overall, the raw acceleration data for all devices differed when data were aggregated based on the mean and RMS for each second, with values generally being more reliable and accurate based on the RMS aggregation method.

A key finding of this study is that the reliability of the raw accelerometry estimates for Apple, Garmin, and Fitbit were similar to ActiGraph. In fact, consumer wearables exhibited

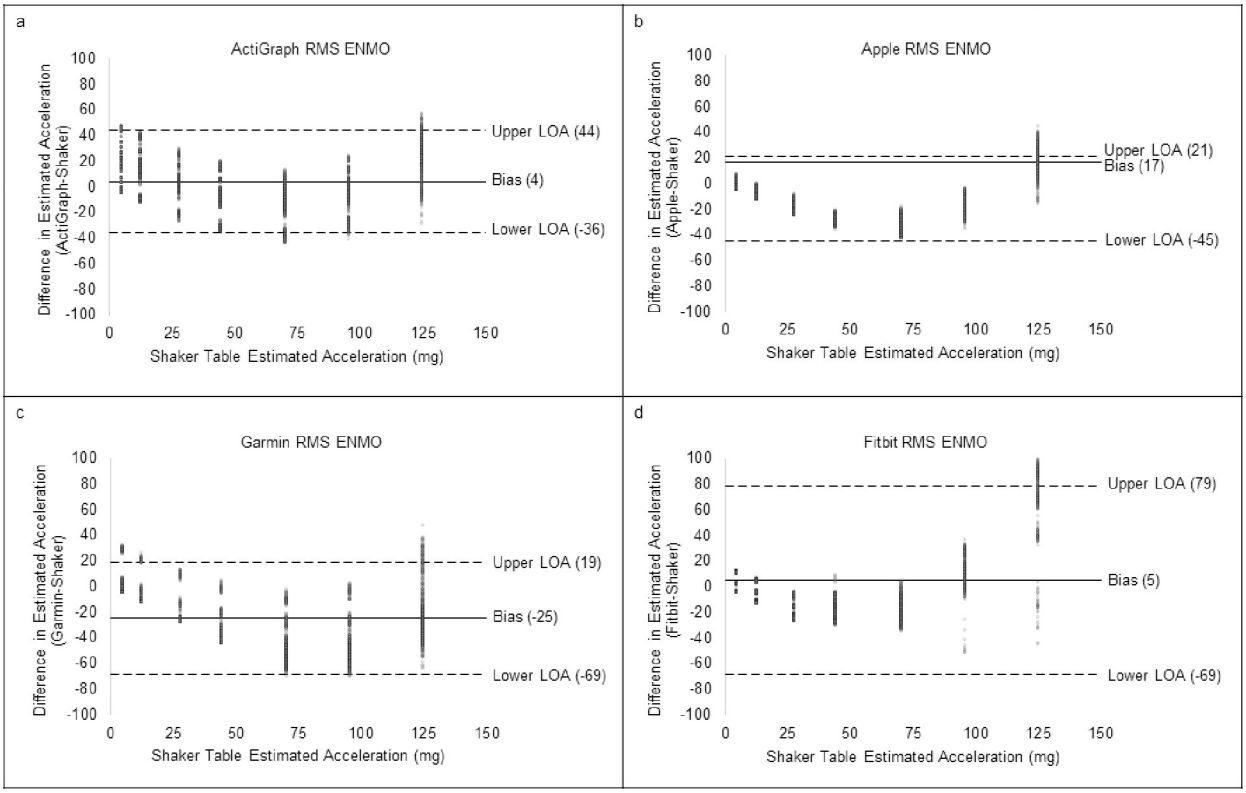

Fig 5. Bland-Altman plots of estimated root mean square ENMO from all devices compared to estimated shaker table acceleration.

https://doi.org/10.1371/journal.pone.0286898.g005

moderate-to-excellent ICC values, with Apple demonstrating nearly perfect reliability (ICC of 0.99). These findings are similar to other studies evaluating the between-device reliability of research-grade devices using a mechanical shaker table. For instance, Powell et al. [39] reported an ICC of 0.99 between 23 RT3 accelerometers and Santos-Lozano et al. [19] reported an ICC of 0.97 between 10 ActiGraph GT3X accelerometers. More recently, studies have explored within-device reliability of various accelerometers and have reported ICCs ranging from 0.77 to 1.00 [40, 41]. Thus, ICCs presented in this study suggest that raw acceleration data collected from Apple, Garmin, and Fitbit provide reliable estimates of movement.

Table 3. Equivalence testing for ActiGraph, Apple, Garmin, Fitbit.

		Mean Difference	Lower 90% Bound	Upper 90% Bound	Interpretation
Mean	ActiGraph	-12.9	-15.8	-10.1	not significantly within the equivalence bounds
	Apple	-29.6	-32.4	-26.8	not significantly within the equivalence bounds
	Garmin	-38.5	-40.8	-36.2	not significantly within the equivalence bounds
	Fitbit	-16.2	-20.0	-12.3	not significantly within the equivalence bounds
RMS	ActiGraph	-9.1	-12.1	-6.0	not significantly within the equivalence bounds
	Apple	-20.5	-23.8	-17.2	not significantly within the equivalence bounds
	Garmin	-34.2	-35.5	-33.0	not significantly within the equivalence bounds
	Fitbit	-3.5	-8.3	1.3	not significantly within the equivalence bounds

Equivalence was set at 10% of the shaker speed (6.3mg) and differences were required to be completely within (±) these bounds to be considered equivalent
Abbreviations: "RMS" Root Mean Square, "mg" milligravity

https://doi.org/10.1371/journal.pone.0286898.t003

In the present study, it is also important to note that raw accelerometry estimates collected from Apple and Fitbit exhibited correlation and concordance with the reference metric that was consistent with ActiGraph. On the other hand, raw acceleration data collected from Garmin exhibited less correlation and concordance with the reference metric than ActiGraph. Our findings for Apple and Fitbit correlation are more consistent with a previous study that reported an excellent Pearson correlation ($r = 0.97$) between accelerations produced by GENEA accelerometers and a mechanical shaker table [30]. These findings suggest that raw acceleration data from Apple and Fitbit produce estimates of movement that are similar to raw acceleration data from ActiGraph. However, more information is needed to determine if the raw acceleration data from Garmin can be used to accurately estimate movement. These findings could be due to hardware differences between devices. For example, the dynamic accelerometer range of the ActiGraph is ±8g [42], while the default accelerometer range for Fitbit is ±4g [43]. The dynamic accelerometer range is an estimate of the greatest amount of acceleration that a device can accurately assess, and thus the relatively smaller accelerometer range of Garmin and Fitbit compared to ActiGraph could have led to more error in Garmin and Fitbit estimates at greater frequencies (S2 and S3 Figs). Differences in the raw acceleration output collected from ActiGraph and the consumer wearables could also be due to the post-processing of the raw data, which has been described previously [20].

Further evidence revealed that, compared to the reference metric, raw acceleration estimates from Apple and Fitbit exhibited absolute differences that were similar to the raw acceleration estimates from ActiGraph. On the other hand, raw acceleration estimates from Garmin exhibited larger absolute errors relative to the raw acceleration estimates from ActiGraph. It is also important to note that raw acceleration data from Apple and Garmin underestimated acceleration output by more than 20mg and 30mg, respectively, compared to raw acceleration estimates from ActiGraph. This is concerning for Garmin, since published intensity thresholds derived from ActiGraph data worn on the non-dominant wrist indicate that sedentary thresholds for children (7-11yrs) are under 35.6mg [26, 27]. Based on these intensity thresholds, it would be difficult to distinguish between sedentary and light intensity thresholds for children using raw acceleration output from Garmin. This may suggest that we need to move away from cut-points, especially since a device-agnostic approach may allow for increased comparability of physical activity estimates across time and between consumer wearables and research-grade devices. One way to summarize raw acceleration data in a device-agnostic manner is to generate open-source Monitor-Independent Movement Summary units (MIMS-units) [44]. MIMS-units could increase the standardization of raw data processing from different devices and reduce between-device variability in estimates of movement [44].

Overall, the findings suggest that raw acceleration output from Apple and Fitbit are similar to raw acceleration output from ActiGraph. However, no device estimates were found to be statistically significantly equivalent to accelerations produced by the reference metric. These limitations with accelerometry are well-documented for distinguishing between sedentary and light activity. For instance, a study using 2-regression models to estimate energy expenditure derived from ActiGraph counts in children (7-13yrs) observed mean absolute percent error values that ranged from 32.5% to 39.4% and 14.5% to 42.9% for sedentary and light activities, respectively [45]. A similar study reported that research-grade accelerometers (i.e., ActiGraph, Actical, and AMP-331) tended to overestimate sedentary and light activities in adults [46]. Though most of the evidence on the associations of device-based sedentary behavior and health is based on accelerometers that infer sedentary time from a lack of movement, this can lead to misclassification of low-movement, non-sedentary behaviors as sedentary behaviors [47]. The absolute errors of ActiGraph, Apple, and Fitbit (~20mg) compared to the reference metric suggest that the relatively small window for sedentary behavior (under 35.6mg) may

pose an issue for estimating physical activity outcomes from accelerometry [29]. Therefore, additional metrics (i.e., heart rate) may need to be combined with accelerometry to improve estimates of these outcomes. An advantage of consumer wearables is their ability to collect accelerometry and heart rate data simultaneously. Thus, it may be possible to leverage the raw acceleration and heart rate data from consumer wearables (i.e., Apple and Fitbit) to overcome limitations with accelerometry alone for estimating physical activity outcomes.

There were several strengths of the present study. The first strength is that accelerations produced by a mechanical shaker table served as the reference to assess the reliability and validity of accelerations produced by various accelerometers. This method allowed for a highly controlled, repeatable evaluation of underlying accelerations produced by various accelerometers shaken in orbital motion at known frequencies. Another strength is that the raw accelerations from devices were evaluated, allowing for between-monitor comparisons of accelerations through elimination of proprietary signal processing that has traditionally been used to derive activity counts from research-grade devices [20]. Additionally, this study evaluated the raw accelerations from consumer wearables, addressing concerns about the proprietary signal processing of these devices [48]. By evaluating the raw accelerations for both research-grade and consumer wearable devices, we were able to compare estimates from the devices on the same metric (mg). Lastly, we calculated Lin's CCC, absolute error, mean bias, and equivalence testing to assess the agreement of the raw accelerometry data from research-grade and consumer wearable devices compared to accelerations produced by a mechanical shaker table. This allowed us to evaluate the agreement of the accelerations between proxy and reference, the overall error of the raw acceleration estimates, and the direction of the average error of the estimates from all devices, whereas other studies only used Pearson correlation to assess validity [22, 30].

Pearson correlation merely measures the covariance between two variables, not the agreement or error. Using these statistics, we were also able to compare the validity metrics produced by the raw acceleration estimates from consumer wearables to the validity metrics produced by the raw acceleration estimates from a research-grade device. This provided preliminary evidence for using the raw acceleration output from consumer wearables to estimate physical activity outcomes. However, the raw acceleration output from consumer wearables needs to be evaluated in settings that resemble free-living activities for children.

The limitations of the present study also need to be acknowledged. The first limitation is that there may have been between trial variability in speed across trials that would systematically affect the findings herein. Another limitation may be the technological advances that have occurred in the consumer wearables evaluated during the project. For instance, the Apple Watch Series 8 was released during the project. However, most of the technological advancements between the Apple Watch Series 7 and the Apple Watch Series 8 are centered on the dual-core processor and the addition of a temperature sensor [49], and thus may not impact accelerometer estimates between devices. Yet, information about the hardware of accelerometers used in consumer wearable devices is largely proprietary. Another limitation may be the post-processing of the raw acceleration data for all devices [20]. The post-processing of the raw acceleration data for all devices is proprietary, so the data is not truly raw. It is also unclear why missing data were present across all Fitbit devices except two. This may have been due to software malfunction with the custom Fitbit app (Slog) that was used to leverage the Fitbit Application Programming Interface.

Conclusions

Findings from this study suggest that raw accelerometry data from Apple, Garmin, and Fitbit are reliable and provide estimates of raw accelerometry that are similar to ActiGraph, except

for Garmin. Additionally, no raw accelerometry estimates were statistically significantly equivalent to the reference. Thus, harmonization approaches across devices like MIMs may be necessary if a truly device-agnostic approach is to be adopted. Future studies should explore using device-agnostic and data harmonization approaches for estimating physical activity from raw accelerometry data produced by Apple and Fitbit in settings that resemble free-living activities for children.

Supporting information

S1 Fig. Orbital mechanical shaker used for testing.
(DOCX)

S2 Fig. Absolute error of the raw acceleration data from all devices by speed compared to the accelerations produced by a mechanical shaker table. Error bars represent standard error.
(DOCX)

S3 Fig. Mean bias of the raw acceleration data from all devices by speed compared to the accelerations produced by a mechanical shaker table. Error bars represent standard error.
(DOCX)

S1 File. Analyses collapsed data_v11.
(XLSX)

S2 File. Analyze data all devices STATA code.
(TXT)

S1 Dataset. Apple with shaker speeds_aggregated_v2 –dataset.
(XLSX)

S2 Dataset. Fitbit with shaker speeds_aggregated_v4 –dataset.
(XLSX)

S3 Dataset. Garmin with shaker_speeds_aggregated_v2 –dataset.
(XLSX)

S4 Dataset. Actigraph_aggregated_v3 drop Actigraphid210 –dataset.
(XLSX)

Author Contributions

Conceptualization: James W. White, III, Srihari Nelakuditi, David E. Brown, III, Russell R. Pate, Gregory J. Welk, Massimiliano de Zambotti, Rahul Ghosal, Yuan Wang, Sarah Burkart, Elizabeth L. Adams, Mvs Chandrashekhar, Bridget Armstrong, Michael W. Beets, R. Glenn Weaver.

Data curation: James W. White, III, Nick Tindall, Srihari Nelakuditi, R. Glenn Weaver.

Formal analysis: James W. White, III, Mvs Chandrashekhar, R. Glenn Weaver.

Funding acquisition: Srihari Nelakuditi, David E. Brown, III, Russell R. Pate, Gregory J. Welk, Massimiliano de Zambotti, Yuan Wang, R. Glenn Weaver.

Investigation: James W. White, III.

Methodology: James W. White, III, Olivia L. Finnegan, Srihari Nelakuditi, David E. Brown, III, Russell R. Pate, Gregory J. Welk, Massimiliano de Zambotti, Yuan Wang, Sarah

Burkart, Elizabeth L. Adams, Mvs Chandrashekhar, Bridget Armstrong, Michael W. Beets, R. Glenn Weaver.

Project administration: James W. White, III, R. Glenn Weaver.

Resources: James W. White, III, Srihari Nelakuditi, David E. Brown, III, Russell R. Pate, Gregory J. Welk, Massimiliano de Zambotti, Rahul Ghosal, Sarah Burkart, Elizabeth L. Adams, Mvs Chandrashekhar, Bridget Armstrong, Michael W. Beets, R. Glenn Weaver.

Software: Nick Tindall, Srihari Nelakuditi.

Supervision: Srihari Nelakuditi, David E. Brown, III, Russell R. Pate, Massimiliano de Zambotti, Yuan Wang, R. Glenn Weaver.

Validation: James W. White, III, Olivia L. Finnegan, Russell R. Pate, Rahul Ghosal, Yuan Wang, Sarah Burkart, Elizabeth L. Adams, Bridget Armstrong, Michael W. Beets, R. Glenn Weaver.

Visualization: James W. White, III, Srihari Nelakuditi, David E. Brown, III, Russell R. Pate, Gregory J. Welk, Massimiliano de Zambotti, Rahul Ghosal, Yuan Wang, R. Glenn Weaver.

Writing – original draft: James W. White, III, R. Glenn Weaver.

Writing – review & editing: James W. White, III, Olivia L. Finnegan, Nick Tindall, Srihari Nelakuditi, David E. Brown, III, Russell R. Pate, Gregory J. Welk, Massimiliano de Zambotti, Rahul Ghosal, Yuan Wang, Sarah Burkart, Elizabeth L. Adams, Mvs Chandrashekhar, Bridget Armstrong, Michael W. Beets, R. Glenn Weaver.

References

1. Duncan GE, Sydeman SJ, Perri MG, Limacher MC, Martin AD. Can sedentary adults accurately recall the intensity of their physical activity? Prev Med. 2001; 33(1):18–26. https://doi.org/10.1006/pmed.2001.0847 PMID: 11482992
2. TROIANO RP, BERRIGAN D, DODD KW, MÂSSE LC, TILERT T, MCDOWELL M. Physical Activity in the United States Measured by Accelerometer. Medicine & Science in Sports & Exercise. 2008; 40 (1):181–8. https://doi.org/10.1249/mss.0b013e31815a51b3 PMID: 18091006
3. Kim Y, Beets MW, Welk GJ. Everything you wanted to know about selecting the "right" Actigraph accelerometer cut-points for youth, but. . .: a systematic review. Journal of Science and Medicine in Sport. 2012; 15(4):311–21.
4. Freedson P, Bowles HR, Troiano R, Haskell W. Assessment of physical activity using wearable monitors: recommendations for monitor calibration and use in the field. Medicine and science in sports and exercise. 2012; 44(1 Suppl 1):S1. https://doi.org/10.1249/MSS.0b013e3182399b7e PMID: 22157769
5. Carpenter A, Frontera A. Smart-watches: a potential challenger to the implantable loop recorder? EP Europace. 2016; 18(6):791–3.
6. Hickey AM, Freedson PS. Utility of Consumer Physical Activity Trackers as an Intervention Tool in Cardiovascular Disease Prevention and Treatment. Prog Cardiovasc Dis. 2016; 58(6):613–9. https://doi.org/10.1016/j.pcad.2016.02.006 PMID: 26943981
7. Jia Y, Wang W, Wen D, Liang L, Gao L, Lei J. Perceived user preferences and usability evaluation of mainstream wearable devices for health monitoring. PeerJ. 2018; 6:e5350. https://doi.org/10.7717/peerj.5350 PMID: 30065893
8. Müller J, Hoch AM, Zoller V, Oberhoffer R. Feasibility of Physical Activity Assessment with Wearable Devices in Children Aged 4–10 Years-A Pilot Study. Front Pediatr. 2018; 6:5. https://doi.org/10.3389/fped.2018.00005 PMID: 29435438
9. Garmin. Instinct® Solar 2020 [Available from: https://www.garmin.com/en-US/p/679335].
10. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and Validity of Commercially Available Wearable Devices for Measuring Steps, Energy Expenditure, and Heart Rate: Systematic Review. JMIR Mhealth Uhealth. 2020; 8(9):e18694. https://doi.org/10.2196/18694 PMID: 32897239

11. O'Driscoll R, Turicchi J, Beaulieu K, Scott S, Matu J, Deighton K, et al. How well do activity monitors estimate energy expenditure? A systematic review and meta-analysis of the validity of current technologies. Br J Sports Med. 2020; 54(6):332–40. https://doi.org/10.1136/bjsports-2018-099643 PMID: 30194221

12. Argent R, Hetherington-Rauth M, Stang J, Tarp J, Ortega FB, Molina-Garcia P, et al. Recommendations for Determining the Validity of Consumer Wearables and Smartphones for the Estimation of Energy Expenditure: Expert Statement and Checklist of the INTERLIVE Network. Sports Med. 2022; 52 (8):1817–32. https://doi.org/10.1007/s40279-022-01665-4 PMID: 35260991

13. Feehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Yoo JY, et al. Accuracy of Fitbit Devices: Systematic Review and Narrative Syntheses of Quantitative Data. JMIR Mhealth Uhealth. 2018; 6(8): e10527. https://doi.org/10.2196/10527 PMID: 30093371

14. Strain T, Wijndaele K, Pearce M, Brage S. Considerations for the Use of Consumer-Grade Wearables and Smartphones in Population Surveillance of Physical Activity. Journal for the Measurement of Physical Behaviour. 2022; 5(1):8–14.

15. Åkerberg A, Arwald J, Söderlund A, Lindén M. An Approach to a Novel Device Agnostic Model Illustrating the Relative Change in Physical Behavior Over Time to Support Behavioral Change. Journal of Technology in Behavioral Science. 2022; 7(2):240–51.

16. Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. Scientific reports. 2018; 8 (1):1–10.

17. Terra API. This is it. . . a comprehensive list of wearable data accessible through APIs today 2022 [Available from: https://blog.tryterra.co/comprehensive-list-of-all-the-wearable-data-that-are-available-through-apis-2bcd35a7307f].

18. Keadle SK, Lyden KA, Strath SJ, Staudenmayer JW, Freedson PS. A Framework to Evaluate Devices That Assess Physical Behavior. Exercise and Sport Sciences Reviews. 2019; 47(4):206–14. https://doi.org/10.1249/JES.0000000000000206 PMID: 31524786

19. Santos-Lozano A, Marín P, Torres-Luque G, Ruiz J, Lucia A, Garatachea N. Technical variablity of the GT3X accelerometer. Medical engineering & physics. 2012; 34:787–90.

20. John D, Sasaki J, Staudenmayer J, Mavilia M, Freedson PS. Comparison of raw acceleration from the GENEA and ActiGraph™ GT3X+ activity monitors. Sensors (Basel). 2013; 13(11):14754–63.

21. Halliday D, Resnick R, Walker J. Fundamentals of physics: John Wiley & Sons; 2013.

22. Davoudi A, Wanigatunga AA, Kheirkhahan M, Corbett DB, Mendoza T, Battula M, et al. Accuracy of Samsung Gear S Smartwatch for Activity Recognition: Validation Study. JMIR Mhealth Uhealth. 2019; 7(2):e11270. https://doi.org/10.2196/11270 PMID: 30724739

23. John D, Miller R, Kozey-Keadle S, Caldwell G, Freedson P. Biomechanical examination of the 'plateau phenomenon' in ActiGraph vertical activity counts. Physiol Meas. 2012; 33(2):219–30. https://doi.org/10.1088/0967-3334/33/2/219 PMID: 22260902

24. Lakens D. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social psychological and personality science. 2017; 8(4):355–62. https://doi.org/10.1177/1948550617697177 PMID: 28736600

25. Bakrania K, Yates T, Rowlands AV, Esliger DW, Bunnewell S, Sanders J, et al. Intensity Thresholds on Raw Acceleration Data: Euclidean Norm Minus One (ENMO) and Mean Amplitude Deviation (MAD) Approaches. PLoS One. 2016; 11(10):e0164045. https://doi.org/10.1371/journal.pone.0164045 PMID: 27706241

26. Hildebrand M, Hansen BH, van Hees VT, Ekelund U. Evaluation of raw acceleration sedentary thresholds in children and adults. Scandinavian Journal of Medicine & Science in Sports. 2017; 27(12):1814–23. https://doi.org/10.1111/sms.12795 PMID: 27878845

27. HILDEBRAND M VAN HEES VT, HANSEN BH, EKELUND U. Age Group Comparability of Raw Accelerometer Output from Wrist- and Hip-Worn Monitors. Medicine & Science in Sports & Exercise. 2014; 46(9):1816–24. https://doi.org/10.1249/MSS.0000000000000289 PMID: 24887173

28. van Hees VT, Fang Z, Langford J, Assah F, Mohammad A, da Silva IC, et al. Autocalibration of accelerometer data for free-living physical activity assessment using local gravity and temperature: an evaluation on four continents. J Appl Physiol (1985). 2014; 117(7):738–44. https://doi.org/10.1152/japplphysiol.00421.2014 PMID: 25103964

29. Published cut-points and how to use them in GGIR: GGIR; [Available from: https://cran.r-project.org/web/packages/GGIR/vignettes/CutPoints.html].

30. Esliger DW, Rowlands AV, Hurst TL, Catt M, Murray P, Eston RG. Validation of the GENEA Accelerometer. Med Sci Sports Exerc. 2011; 43(6):1085–93. https://doi.org/10.1249/MSS.0b013e31820513be PMID: 21088628

31. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016; 15(2):155–63. https://doi.org/10.1016/j.jcm.2016.02.012 PMID: 27330520

32. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med. 2018; 18(3):91–3. https://doi.org/10.1016/j.tjem.2018.08.001 PMID: 30191186

33. Dancey CP, Reidy J. Statistics without maths for psychology: Pearson education; 2007.

34. Altman DG. Practical statistics for medical research Chapman and Hall. London and New York. 1991.

35. Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of pharmacokinetics and biopharmaceutics. 1987; 15(6):657–80. https://doi.org/10.1007/BF01068419 PMID: 3450848

36. Rogers JL, Howard KI, Vessey JT. Using significance tests to evaluate equivalence between two experimental groups. Psychological bulletin. 1993; 113(3):553. https://doi.org/10.1037/0033-2909.113.3.553 PMID: 8316613

37. Dixon PM, Saint-Maurice PF, Kim Y, Hibbing P, Bai Y, Welk GJ. A primer on the use of equivalence testing for evaluating measurement agreement. Medicine and science in sports and exercise. 2018; 50 (4):837. https://doi.org/10.1249/MSS.0000000000001481 PMID: 29135817

38. Chowdhury EA, Western MJ, Nightingale TE, Peacock OJ, Thompson D. Assessment of laboratory and daily energy expenditure estimates from consumer multi-sensor physical activity monitors. PLoS One. 2017; 12(2):e0171720. https://doi.org/10.1371/journal.pone.0171720 PMID: 28234979

39. POWELL SM, JONES DI, ROWLANDS AV. Technical Variability of the RT3 Accelerometer. Medicine & Science in Sports & Exercise. 2003; 35(10):1773–8. https://doi.org/10.1249/01.MSS.0000089341.68754.BA PMID: 14523319

40. Nicolella DP, Torres-Ronda L, Saylor KJ, Schelling X. Validity and reliability of an accelerometer-based player tracking device. PLoS One. 2018; 13(2):e0191823. https://doi.org/10.1371/journal.pone.0191823 PMID: 29420555

41. Vanhelst J, Fardy PS, Beghin L. Technical variability of the Vivago® wrist-worn accelerometer. J Sports Sci. 2014; 32(19):1768–74.

42. ActiGraph; [cited 2023 03/10/2023]. Available from: https://actigraphcorp.com/actigraph-wgt3x-bt/.

43. Isakeit T. Fitbit Sense Teardown 2021 [Available from: https://www.ifixit.com/Teardown/Fitbit+Sense+Teardown/137130].

44. John D, Tang Q, Albinali F, Intille S. An Open-Source Monitor-Independent Movement Summary for Accelerometer Data Processing. J Meas Phys Behav. 2019; 2(4):268–81. https://doi.org/10.1123/jmpb.2018-0068 PMID: 34308270

45. Kim Y, Crouter SE, Lee JM, Dixon PM, Gaesser GA, Welk GJ. Comparisons of prediction equations for estimating energy expenditure in youth. J Sci Med Sport. 2016; 19(1):35–40. https://doi.org/10.1016/j.jsams.2014.10.002 PMID: 25459235

46. Crouter SE, Churilla JR, Bassett DR Jr. Estimating energy expenditure using accelerometers. Eur J Appl Physiol. 2006; 98(6):601–12. https://doi.org/10.1007/s00421-006-0307-5 PMID: 17058102

47. Rowlands AV, Olds TS, Hillsdon M, Pulsford R, Hurst TL, Eston RG, et al. Assessing sedentary behavior with the GENEActiv: introducing the sedentary sphere. Med Sci Sports Exerc. 2014; 46(6):1235–47. https://doi.org/10.1249/MSS.0000000000000224 PMID: 24263980

48. Shei RJ, Holder IG, Oumsang AS, Paris BA, Paris HL. Wearable activity trackers-advanced technology or advanced marketing? Eur J Appl Physiol. 2022; 122(9):1975–90. https://doi.org/10.1007/s00421-022-04951-1 PMID: 35445837

49. Apple Watch models: Apple; [Available from: https://www.apple.com/watch/compare/].