

1

2

3

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

6

7 James W. White III^{1*}, Olivia Finnegan^{1¶}, Nick Tindall^{1¶}, Srihari Nelakuditi^{1¶}, David E. Brown
8 III^{1,4¶}, Russ Pate Ph.D.^{1¶}, Gregory J. Welk^{2¶}, Massimiliano de Zambotti^{3¶}, Rahul Ghosal^{1¶}, Yuan
9 Wang^{1¶}, Sarah Burkart^{1¶}, Elizabeth L. Adams^{1¶}, Mvs Chandrashekhar^{1¶}, Bridget Armstrong^{1¶},
10 Michael W. Beets^{1¶}, R. Glenn Weaver^{1¶}

11

12 ¹ Department of Exercise Science, University of South Carolina, Columbia, South Carolina, USA

13 ² Kinesiology, Iowa State University, Ames, IA, USA

14 ³ SRI International, Menlo Park, CA, USA

15 ⁴ Division of Pediatric Pulmonology, Pediatric Sleep Medicine, Prisma Health Richland Hospital,
16 Columbia, South Carolina, USA

17

18 #a921 Assembly Street, PHRC 115: Department of Exercise Science, University of South
19 Carolina, Columbia, South Carolina, USA

20

21 #b921 Assembly Street, PHRC 115: Department of Exercise Science, University of South
22 Carolina, Columbia, South Carolina, USA

23

24

25 James W. White III

26 E-mail: Jww4@email.sc.edu (JW)

27 [¶]These authors contributed equally to this work.

28 Acknowledgements. None

30 **Abstract**

31 The purpose of this study was to evaluate the reliability and validity of the raw
32 accelerometry output from research-grade and consumer wearable devices compared to
33 accelerations produced by a mechanical shaker table. Raw accelerometry data from a total of 40
34 devices (i.e., n=10 ActiGraph wGT3X-BT, n=10 Apple Watch Series 7, n=10 Garmin
35 Vivoactive 4S, and n=10 Fitbit Sense) were compared to the criterion accelerations produced by
36 an orbital shaker table at speeds ranging from 0.6 Hz (4.4 milligravity-mg) to 3.2 Hz (124.7mg).
37 For reliability testing, identical devices were oscillated at 0.6 and 3.2 Hz for 5 trials that lasted 2
38 minutes each. For validity testing, devices were oscillated for 1 trial across 7 speeds that lasted 2
39 minutes each. The intraclass correlation coefficient (ICC) was calculated to test inter-device
40 reliability. Pearson product moment, Lin's concordance correlation coefficient (CCC), absolute
41 error, and mean bias were calculated to assess the validity between the raw estimates from the
42 devices and the criterion metric. Estimates produced by the raw accelerometry data from Apple
43 and ActiGraph were more reliable ICCs=0.99 and 0.97 than Garmin and Fitbit ICCs=0.88 and
44 0.88, respectively. Estimates from ActiGraph, Apple, and Fitbit devices exhibited excellent
45 concordance with the criterion CCCs=0.88, 0.83, and 0.85, respectively, while estimates from
46 Garmin exhibited moderate concordance CCC=0.59 based on the mean aggregation method.
47 ActiGraph, Apple, and Fitbit produced similar absolute errors=16.9mg, 21.6mg, and 22.0mg,
48 respectively, while Garmin produced higher absolute error=32.5mg compared to the criterion
49 based on the mean aggregation method. ActiGraph produced the lowest mean bias 0.0mg
50 (95%CI=-40.0, 41.0) based on the mean aggregation method. Raw accelerometry data collected
51 from Apple and Fitbit are comparable to ActiGraph. However, raw accelerometry data from

- 52 Garmin appears to be different. Future studies may be able to develop algorithms using device-
- 53 agnostic methods for estimating physical activity from consumer wearables.

54 **Introduction**

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

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

71 However, these studies are limited because they rely on estimates of physical activity that
72 are derived from proprietary algorithms developed by the companies that produce these devices
73 (e.g., Apple, Garmin, Fitbit, etc.). This is a key limitation because these algorithms are unavailable
74 for review by researchers [12]. The drawbacks of estimating physical activity based on proprietary
75 algorithms are that it is unclear whether best practice recommendations were used to develop these

76 algorithms, and the algorithms could be updated by these companies at any time unbeknownst to
77 the user. Thus, estimates of physical activity collected from the same device across time may
78 provide different estimates of activity due solely to changes in the underlying algorithms that
79 produce these metrics.

80 An alternative, device-agnostic or monitor-independent approach may address these
81 limitations by enabling data from any device to be processed using the same algorithm or
82 processing methodology [13, 14]. A device-agnostic approach is a realistic possibility as consumer
83 wearables have released application programming interfaces (API) that allow access to the raw
84 accelerometry data (i.e., x, y, z axis readings collected by these devices [15]. This has the potential
85 to increase the comparability of physical activity estimates across time and between different
86 consumer wearables and research-grade devices.

87 A necessary first step to applying a device-agnostic approach to raw accelerometry data
88 collected by consumer wearables is to conduct calibration studies that explore the reliability and
89 validity of the underlying acceleration output produced by these devices [16]. This testing will
90 provide insight into the reliability and validity of the raw acceleration output from consumer
91 wearables in a controlled environment, prior to evaluating how human variation impacts the raw
92 acceleration estimates from these devices [16]. Therefore, this study will evaluate the between-
93 device reliability and validity of the raw acceleration output from research-grade and consumer
94 wearable devices, compared to accelerations produced by a mechanical shaker table at various
95 speeds as the criterion measure. It is important to include research-grade devices in this study
96 because it allows us to evaluate if the raw accelerometry estimates from consumer wearables are
97 comparable to the raw accelerometry estimates of research-grade devices when compared to more
98 direct estimates of acceleration from a mechanical shaker table. While studies have previously

99 examined the ActiGraph with this methodology [17, 18], this is the first study that we are aware
100 of to report shaker table outcomes with consumer-grade devices.

101 **Methods**

102 Raw accelerometry data from a total of 40 devices were evaluated in this study. The
103 research-grade devices included n=10 ActiGraph wGT3X-BT (ActiGraph; ActiGraph LLC
104 Pensacola, FL). The consumer wearable devices included n=10 Apple Watch Series 7 (Apple;
105 Apple Technology Company, Cupertino, CA), n=10 Garmin Vioactive 4S (Garmin; Garmin Ltd.,
106 Olathe, KS), and n=10 Fitbit Sense (Fitbit; Google LLC, San Francisco, CA). Inter-device
107 reliability and validity of raw accelerations for all devices were tested, with accelerations produced
108 by a mechanical shaker table (Scientific Industries, Bohemia, NY; Mini-300 Orbital-Genie, Model
109 1500) as the criterion. Each device was securely mounted directly to the twin ratcheting clamps of
110 a mechanical shaker table (Fig 1) that produces controlled oscillations at frequencies between
111 approximately $f_{shaker}=0.6$ and 5 Hertz (Hz). We converted f_{shaker} in Hz to acceleration using the
112 expression for centripetal acceleration, $a_{orbital} = v^2/r_{orbital}$ [19], where $r_{orbital}$ is the radius of
113 rotation for the orbital shaker $r_{orbital}$. From the manual for this particular shaker (supplementary
114 https://cdn.shopify.com/s/files/1/0489/6990/8374/files/SI-M1600_Manual.pdf?v=1617998279),
115 the specified diameter of the orbit is $2r_{orbital}=1.9\text{cm}$ and the rotational speed is given by $v = 2\pi$
116 $r_{orbital}f_{shaker}$, since for each complete cycle of 2π radians, the table traverses a distance of
117 circumference $2\pi r_{orbital}$ in time $1/f_{shaker}$. In other words:

$$118 \quad a_{orbital}(\text{cm}/\text{s}^2) = 4\pi^2 r_{orbital} f_{shaker}^2$$

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

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

124 **Figure 1. Orbital mechanical shaker used for shaker testing.**

125 **Device software**

126 ActiGraphs were initialized to provide output from each directional axis using ActiLife
127 software (version 6.13.4; ActiGraph LLC, Pensacola, FL). Garmin devices were initialized, and
128 data were recorded in RawLogger (version 1.0.20211201a) and exported through Garmin Connect
129 softwareTM. Apple devices were initialized, and data were recorded in SensorLog (version 5.2) and
130 exported into comma-separated values (CSV) files via Health Auto Export (version 6.3).
131 RawLogger and SensorLog are user-written apps that leverage the device-specific Application
132 Programming Interface (API) to collect the underlying sensor data on the respective devices.
133 RawLogger is available for download through the Connect IQTM store on the Garmin ConnectTM
134 app, and SensorLog and Health Auto Export are available for download through the App Store.
135 The research team developed a custom Fitbit app (Slog) leveraging the Fitbit API for the same
136 purpose, and Fitbit devices were initialized, and data were recorded and exported through this app.
137 The GitHub code for the custom Fitbit app is available at
138 <https://github.com/ntindallUSC/Slog/tree/main>. Sampling frequencies from 25 Hz to 100 Hz were
139 recorded based on the capabilities of the ActiGraph (100 Hz), Apple (100 Hz), Garmin (25 Hz),
140 and Fitbit (50 Hz).

141 **Reliability testing**

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

151 **Validity testing**

152 For validity testing, five identical devices were mounted side-by-side until all devices were
153 run through the validity trials. The trials lasted 14 minutes and 30 seconds. Consistent with past
154 validation studies [18, 20], each trial began with a 15-second rest period (i.e., no movement)
155 followed by a standardized series of oscillations at seven frequencies (i.e., 3.2 Hz, 2.8 Hz, 2.4 Hz,
156 1.9 Hz, 1.5 Hz, 1.0 Hz, 0.6 Hz) lasting two minutes each. These frequencies were chosen because
157 they are consistent with human movement ranging from 1.5 to 16 mph [21]. The start and stop
158 times were noted at each frequency for both research-grade and consumer wearable devices. Each
159 trial ended with another 15-second rest period. A minimum of 2 trials were conducted for each
160 device brand, totaling 8 trials. Trials/devices with missing data due to device malfunction: Apple
161 (n=4) and Fitbit (n=1) or shaker table malfunction (n=1) were repeated to minimize missing data;
162 no trials had to be repeated for ActiGraph or Garmin devices. Following all testing, raw
163 acceleration data for both research-grade and consumer wearable devices were downloaded and

164 converted to a CSV file using ActiLife software and the device-specific user-written apps,
165 respectively.

166 **Data processing**

167 Raw acceleration data from all devices (i.e., ActiGraph, Apple, Garmin, and Fitbit) were
168 extracted from the middle minute of each 2-minute oscillation frequency. Consistent with past
169 research, Euclidean Norm Minus One (ENMO) was calculated [4, 5]. All values were multiplied
170 by 1000 (milligravity-mg) to be consistent with published intensity thresholds based on the GGIR
171 package for accelerometry in R statistical software [22]. Data were aggregated to the second level
172 by extracting the mean and root mean square (RMS) value for each second for all devices for
173 ENMO. Both mean and RMS were calculated as both methods have been calculated previously,
174 which suggests that there is no consensus on how raw accelerometry data should be aggregated
175 [18, 20, 23].

176 **Correlation coefficients**

177 To test reliability, a single, absolute intraclass correlation coefficient (ICC) was calculated
178 for all devices. ICC values less than 0.50 were defined as poor reliability, between 0.50 and 0.75
179 as moderate reliability, between 0.75 and 0.90 as good reliability, and greater than 0.90 as excellent
180 reliability [24]. Prior to statistical analyses for validity testing, descriptive means and standard
181 deviations for the mean and RMS were calculated across devices for each speed ranging from 0.6
182 to 3.2 Hz. For the validity testing, Pearson product moment (r) and Lin's concordance correlation
183 coefficient (CCC) were calculated to assess correlation and agreement of raw acceleration data
184 from ActiGraph and consumer wearable devices compared to the criterion (i.e., acceleration from
185 the shaker table) [25]. Pearson product moment interpretations were defined based on Dancey and

186 Reidy [26], and Lin's concordance correlation coefficient was defined similarly based on
187 recommendations from Altman (1991), with coefficients less than 0.20 as poor and greater than
188 0.80 as excellent [27].

189 **Discrepancy analyses**

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

199 **Results**

200 For reliability, ICCs (95% confidence intervals) are presented for the raw acceleration data
201 from all devices for both aggregation methods (i.e., mean and RMS) for all devices in Table 1.
202 The ICCs for ActiGraph were 0.97 (0.92, 0.99) and 0.97 (0.93, 0.98) for the mean and RMS
203 aggregation methods, respectively. The ICCs for Apple were 0.99 (0.99, 0.99) and 0.99 (0.99,
204 1.00) for the mean and RMS, respectively. The ICCs for Garmin were 0.88 (0.82, 0.92) and 0.90
205 (0.85, 0.93) for the mean and RMS aggregation methods, respectively. The ICCs for Fitbit were
206 0.88 (0.86, 0.89) and 0.87 (0.85, 0.88) for the mean and RMS aggregation methods, respectively.

207 **Table 1. Summary of Intraclass Correlation Coefficients for All Devices Aggregated based**
 208 **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)

209 ^a95CI = 95% confidence interval; RMS = root mean square

210 For validity, a summary table of outcomes based on the raw acceleration data from all devices is
 211 presented in Table 2. Fig 2 shows the concordance of the raw acceleration data from all devices
 212 compared to the criterion metric. Fig 3 shows the absolute error of the raw acceleration data from
 213 all devices compared to the criterion metric. Fig 4 shows the mean bias of the raw acceleration
 214 data from all devices compared to the criterion metric.

Table 2. Summary of Validity Outcomes for All Devices Aggregated based on the Mean and Root Mean Square.

	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

215 ^aSD = standard deviation; mg = milligravity

216 **Fig 2. Lin's Concordance Correlation Coefficient of the Raw Acceleration Data from all**
 217 **Devices Compared to the Accelerations Produced by a Mechanical Shaker Table.**

218 **Fig 3. Absolute Error of the Raw Acceleration Data from all Devices Compared to the**
 219 **Accelerations Produced by a Mechanical Shaker Table.**

220 **Fig 4. Mean Bias of the Raw Acceleration Data from all Devices Compared to the**
 221 **Accelerations Produced by a Mechanical Shaker Table.**

222 Pearson product moment correlations between raw accelerometry estimates for ActiGraph
 223 and the criterion metric were $r=0.88$ and $r=0.89$ for the mean and RMS aggregation methods,
 224 respectively. CCCs (95% confidence intervals) when compared to the shaker table were $r_c=0.88$
 225 (0.87, 0.80) and $r_c=0.88$ (0.88, 0.89) for the mean and RMS aggregation methods, respectively.
 226 Mean bias (95% confidence intervals) was 0.0mg (-40.0, 41.0) and 4.0mg (-36.0, 44.0), and
 227 absolute error was 16.9mg and 16.7mg for the mean and RMS aggregation methods, respectively.

228 Pearson product moment correlations between raw accelerometry estimates for Apple and
 229 the criterion metric were $r=0.94$ and $r=0.94$ for the mean and RMS aggregation methods,
 230 respectively. CCCs when compared to the shaker table were $r_c=0.83$ (0.82, 0.83) and $r_c=0.90$ (0.89,
 231 0.90) for the mean and RMS aggregation methods, respectively. Mean bias (95% confidence

232 intervals) was -21.0mg (-50.0, 7.0) and -12.0mg (-45.0, 21.0), and absolute error was 21.6mg and
233 18.0mg for the mean and RMS aggregation methods, respectively.

234 Pearson product moment correlations between raw accelerometry estimates for Garmin and
235 the criterion metric were $r=0.79$ and $r=0.84$ for the mean and RMS aggregation methods,
236 respectively. CCCs when compared to the shaker table were $r_c=0.59$ (0.58, 0.60) and $r_c=0.70$ (0.69,
237 0.71) for the mean and RMS aggregation methods, respectively. Mean bias (95% confidence
238 intervals) was -30.0mg (-80.0, 19.0) and -25.0mg (-69.0, 19.0), and absolute error was 32.5mg and
239 28.1mg for the mean and RMS aggregation methods, respectively.

240 Pearson product moment correlations between raw accelerometry estimates for Fitbit and
241 the criterion metric were $r=0.91$ and $r=0.92$ for the mean and RMS aggregation methods,
242 respectively. CCCs when compared to the shaker table were $r_c=0.85$ (0.84, 0.86) and $r_c=0.79$
243 (0.78,0.80) for the mean and RMS aggregation methods, respectively. Mean bias (95% confidence
244 intervals) was -8.0mg and 5.0mg, and absolute error was 22.0mg and 24.2mg for the mean and
245 RMS aggregation methods, respectively.

246 **Discussion**

247 The aim of this study was to evaluate the between-device reliability and validity of the raw
248 acceleration output from research-grade (i.e., ActiGraph wGT3X-BT) and consumer wearable
249 devices (i.e., Apple Watch Series 7, Garmin Vivoactive 4S, and Fitbit Sense) compared to
250 accelerations produced by a mechanical shaker table. The raw acceleration data collected from all
251 devices exhibited good-to-excellent between-device reliability based on the mean and RMS
252 aggregation methods. For validity, the raw acceleration data from all devices exhibited a strong
253 positive correlation to the criterion metric with moderate-to-excellent concordance no matter the

254 aggregation method. Except for Garmin, the raw acceleration data collected from consumer
255 wearables demonstrated absolute errors that were consistent with ActiGraph. Moreover, the raw
256 acceleration data collected from consumer wearables underestimated acceleration output to a
257 greater degree than ActiGraph when compared to the accelerations produced by the mechanical
258 shaker table. Overall, the raw acceleration data for all devices differed when data were aggregated
259 based on the mean and RMS for each second, with values generally being more reliable and
260 accurate based on the RMS aggregation method.

261 A key finding of this study is that the reliability for Apple, Garmin, and Fitbit was similar
262 to ActiGraph. In fact, consumer wearables exhibited moderate-to-excellent ICC values, with Apple
263 demonstrating nearly perfect reliability with an ICC of 0.99. These findings are similar to other
264 studies evaluating the between-device reliability of research-grade devices using a mechanical
265 shaker table. For instance, Powell et al. [28] reported an ICC of 0.99 between 23 RT3
266 accelerometers and Santos-Lozano et al. [17] reported an ICC of 0.97 between 10 ActiGraph
267 GT3X accelerometers. More recently, studies have explored within-device reliability of various
268 accelerometers and have reported ICCs ranging from 0.77 to 1.00 [29, 30]. Thus, ICCs based on
269 the raw acceleration data collected from consumer wearables in the present study support their use
270 as a reliable tool to assess physical activity.

271 In the present study, it is also important to note that raw accelerometry estimates collected
272 from Apple and Fitbit exhibited correlation and concordance with the criterion metric that was
273 consistent with ActiGraph. On the other hand, raw acceleration data collected from Garmin
274 exhibited less correlation and concordance with the criterion metric than ActiGraph. Our findings
275 for Apple and Fitbit correlation are more consistent with a previous study that reported an excellent
276 Pearson correlation ($r=0.97$) for accelerations produced by GENE A accelerometers and a

277 mechanical shaker table [23]. These findings suggest that raw acceleration data from Apple and
278 Fitbit may produce comparable estimates of activity than raw acceleration data from ActiGraph.
279 More information is needed to determine whether the raw acceleration data from Garmin could be
280 used to accurately estimate physical activity. These findings could be due to the hardware
281 differences between devices. For example, the dynamic accelerometer range of the ActiGraph is
282 $\pm 8g$ [31], while the default accelerometer range for Fitbit is $\pm 4g$ [32]. The dynamic accelerometer
283 range is an estimate of the greatest amount of acceleration that a device can accurately assess, and
284 thus the relatively smaller accelerometer range of Garmin and Fitbit compared to ActiGraph could
285 have led to more error in Garmin and Fitbit raw accelerometry estimates at greater frequencies (S
286 Fig 1 and 2). Differences in the raw acceleration output collected from ActiGraph and the
287 consumer wearables could also be due to the post-processing of the raw data, which has been
288 described previously [18].

289 Further evidence revealed that, compared to the criterion metric, raw acceleration estimates
290 from Apple and Fitbit exhibited absolute errors similar to the raw acceleration estimates from
291 ActiGraph, while raw acceleration estimates from Garmin exhibited larger absolute errors relative
292 to the raw acceleration estimates from ActiGraph. It is also important to note that raw acceleration
293 data from Apple and Garmin underestimated acceleration output by more than 20mg and 30mg,
294 respectively, compared to raw acceleration estimates from ActiGraph. This evidence is concerning
295 for Garmin, considering that published intensity thresholds derived from ActiGraph worn on the
296 non-dominant wrist indicates that sedentary thresholds for children (7-11yrs) are under 35.6mg
297 [33, 34]. Based on these intensity thresholds, it would be difficult to distinguish between sedentary
298 and light intensity thresholds for children using raw acceleration output from Garmin. This may
299 suggest that we need to move away from cut-points, especially since a device-agnostic approach

300 may allow for increased comparability of physical activity estimates across time and between
301 consumer wearables and research-grade devices. However, more work is needed, specifically with
302 Garmin. A device-agnostic approach using raw accelerometry data from Garmin could lead to
303 different estimates of activity because the raw accelerometry output is different from ActiGraph,
304 Apple, and Fitbit.

305 Overall, the findings suggest that the raw acceleration output from Apple and Fitbit is
306 comparable to the raw acceleration output from ActiGraph. However, limitations with
307 accelerometry are well-documented for distinguishing between sedentary and light activity. For
308 instance, a study using 2-regression models to estimate energy expenditure derived from
309 ActiGraph counts observed mean absolute percent error values that ranged from 32.5% to 39.4%
310 and 14.5% to 42.9% for sedentary and light activities, respectively, in children 7-13yrs [35]. A
311 similar study reported that research-grade accelerometers (i.e., ActiGraph, Actical, and AMP-331)
312 tended to overestimate sedentary and light activities in adults [36]. Though most of the evidence
313 on the associations of objectively assessed sedentary behavior and health is based on
314 accelerometers that infer sedentary time from a lack of movement, this can lead to misclassification
315 of low-movement, non-sedentary behaviors as sedentary behaviors [37]. The absolute errors of
316 ActiGraph, Apple, and Fitbit (~20mg) compared to the criterion metrics suggest that the relatively
317 small window for sedentary behavior (under 35.6mg) may pose an issue for estimating physical
318 activity outcomes from accelerometry [22]. Therefore, additional metrics (i.e., heart rate) may need
319 to be combined with accelerometry to improve estimates of these outcomes. An advantage of
320 consumer wearables is their ability to collect acceleration and heart rate data simultaneously. Thus,
321 it may be possible to leverage the raw acceleration and heart rate data from consumer wearables

322 (i.e., Apple and Fitbit) to overcome limitations with accelerometry alone for estimating physical
323 activity outcomes.

324 There were several strengths of the present study. The first strength is that accelerations
325 produced by a mechanical shaker table served as the criterion to assess the reliability and validity
326 of accelerations produced by various accelerometers. This method allowed for a highly controlled,
327 repeatable evaluation of underlying accelerations produced by various accelerometers shaken in
328 orbital motion at known frequencies. Another strength is that the raw accelerations from devices
329 were evaluated, allowing for between-monitor comparisons of accelerations through elimination
330 of proprietary signal processing that has traditionally been used to derive activity counts from
331 research-grade devices [18]. Additionally, this study evaluated the raw accelerations from
332 consumer wearables, addressing concerns about the proprietary signal processing of these devices
333 [38]. By evaluating the raw accelerations for both research-grade and consumer wearable devices,
334 we were able to compare acceleration estimates from the devices based on the same metric (mg).
335 Lastly, we calculated Lin's CCC, absolute error, and mean bias to assess the agreement of the raw
336 accelerometry data from research-grade and consumer wearable devices compared to accelerations
337 produced by the shaker table. This allowed us to evaluate the agreement of the accelerations
338 between proxy and criterion, the overall error of the raw acceleration estimates, and the direction
339 of the average error of the raw acceleration estimates from all devices, whereas other studies used
340 Pearson correlation to assess validity [20, 23].

341 Pearson correlation merely measures the covariance between two variables, not the
342 agreement or error. Using these statistics, we were also able to compare the validity metrics
343 produced by the raw acceleration estimates from consumer wearables to the validity metrics
344 produced by the raw acceleration estimates from a research-grade device. This provided

345 preliminary evidence for using the raw acceleration output of consumer wearables to estimate
346 physical activity outcomes. However, the raw acceleration output from consumer wearables needs
347 to be evaluated in settings that resemble free-living activities for children.

348 The limitations of the present study also need to be acknowledged. One limitation may be
349 the technological advances that have occurred in the consumer wearables evaluated during the
350 project. For instance, the Apple Watch Series 8 was released during the project. However, most of
351 the technological advancements between the Apple Watch Series 7 and the Apple Watch Series 8
352 are centered on the dual-core processor and the addition of a temperature sensor [39], and thus
353 may not impact accelerometer estimates between devices. Yet, information about the hardware of
354 accelerometers used in consumer wearable devices is largely proprietary. Another limitation may
355 be the post-processing of the raw acceleration data for all devices [18]. The post-processing of the
356 raw acceleration data for all devices is proprietary, so the acceleration data is not truly raw. It is
357 also unclear why missing data were present across all Fitbit devices except two. This may have
358 been due to software malfunction with the custom Fitbit app (Slog) that was used to leverage the
359 Fitbit Application Programming Interface.

360 **Conclusions**

361 Findings from this study suggest that raw accelerometry data from Apple, Garmin, and
362 Fitbit are reliable and provide estimates of raw accelerometry that are similar to ActiGraph.
363 Additionally, raw accelerometry estimates for Apple and Fitbit are comparable to raw
364 accelerometry estimates from ActiGraph, while raw accelerometry estimates from Garmin differ
365 from estimates from ActiGraph. Yet, limitations with accelerometry are well-documented for
366 distinguishing between sedentary and light activity. Consumer wearables' ability to capture both

367 accelerometry and heart rate could improve estimates of activity, especially sedentary and light
368 activity. Future studies should explore using a device-agnostic approach for estimating physical
369 activity from raw accelerometry data produced by Apple and Fitbit in settings that resemble free-
370 living activities for children.

371 **Acknowledgements**

372 None.

373 **References**

- 374 1. Duncan GE, Sydeman SJ, Perri MG, Limacher MC, Martin AD. Can sedentary adults
375 accurately recall the intensity of their physical activity? *Prev Med.* 2001;33(1):18-26.
- 376 2. TROIANO RP, BERRIGAN D, DODD KW, MÂSSE LC, TILERT T, MCDOWELL M.
377 Physical Activity in the United States Measured by Accelerometer. *Medicine & Science in*
378 *Sports & Exercise.* 2008;40(1):181-8.
- 379 3. Kim Y, Beets MW, Welk GJ. Everything you wanted to know about selecting the “right”
380 Actigraph accelerometer cut-points for youth, but...: a systematic review. *Journal of Science and*
381 *Medicine in Sport.* 2012;15(4):311-21.
- 382 4. Freedson P, Bowles HR, Troiano R, Haskell W. Assessment of physical activity using
383 wearable monitors: recommendations for monitor calibration and use in the field. *Medicine and*
384 *science in sports and exercise.* 2012;44(1 Suppl 1):S1.
- 385 5. Carpenter A, Frontera A. Smart-watches: a potential challenger to the implantable loop
386 recorder? *EP Europace.* 2016;18(6):791-3.
- 387 6. Hickey AM, Freedson PS. Utility of Consumer Physical Activity Trackers as an
388 Intervention Tool in Cardiovascular Disease Prevention and Treatment. *Prog Cardiovasc Dis.*
389 2016;58(6):613-9.
- 390 7. Jia Y, Wang W, Wen D, Liang L, Gao L, Lei J. Perceived user preferences and usability
391 evaluation of mainstream wearable devices for health monitoring. *PeerJ.* 2018;6:e5350.
- 392 8. Müller J, Hoch AM, Zoller V, Oberhoffer R. Feasibility of Physical Activity Assessment
393 with Wearable Devices in Children Aged 4-10 Years-A Pilot Study. *Front Pediatr.* 2018;6:5.
- 394 9. Garmin. Instinct® Solar 2020 [Available from: [https://www.garmin.com/en-](https://www.garmin.com/en-US/p/679335)
395 [US/p/679335](https://www.garmin.com/en-US/p/679335)].
- 396 10. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and
397 Validity of Commercially Available Wearable Devices for Measuring Steps, Energy
398 Expenditure, and Heart Rate: Systematic Review. *JMIR Mhealth Uhealth.* 2020;8(9):e18694.
- 399 11. O'Driscoll R, Turicchi J, Beaulieu K, Scott S, Matu J, Deighton K, et al. How well do
400 activity monitors estimate energy expenditure? A systematic review and meta-analysis of the
401 validity of current technologies. *Br J Sports Med.* 2020;54(6):332-40.
- 402 12. Argent R, Hetherington-Rauth M, Stang J, Tarp J, Ortega FB, Molina-Garcia P, et al.
403 Recommendations for Determining the Validity of Consumer Wearables and Smartphones for
404 the Estimation of Energy Expenditure: Expert Statement and Checklist of the INTERLIVE
405 Network. *Sports Med.* 2022;52(8):1817-32.
- 406 13. Åkerberg A, Arwald J, Söderlund A, Lindén M. An Approach to a Novel Device
407 Agnostic Model Illustrating the Relative Change in Physical Behavior Over Time to Support
408 Behavioral Change. *Journal of Technology in Behavioral Science.* 2022;7(2):240-51.
- 409 14. Willetts M, Hollowell S, Aslett L, Holmes C, Doherty A. Statistical machine learning of
410 sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants.
411 *Scientific reports.* 2018;8(1):1-10.
- 412 15. Terra API. This is it... a comprehensive list of wearable data accessible through APIs
413 today. 2022.
- 414 16. Keadle SK, Lyden KA, Strath SJ, Staudenmayer JW, Freedson PS. A Framework to
415 Evaluate Devices That Assess Physical Behavior. *Exercise and Sport Sciences Reviews.*
416 2019;47(4):206-14.

- 417 17. Santos-Lozano A, Marín P, Torres-Luque G, Ruiz J, Lucia A, Garatachea N. Technical
418 variability of the GT3X accelerometer. *Medical engineering & physics*. 2012;34:787-90.
- 419 18. John D, Sasaki J, Staudenmayer J, Mavilia M, Freedson PS. Comparison of raw
420 acceleration from the GENEActiv and ActiGraph™ GT3X+ activity monitors. *Sensors (Basel)*.
421 2013;13(11):14754-63.
- 422 19. Halliday D, Resnick R, Walker J. *Fundamentals of physics*: John Wiley & Sons; 2013.
- 423 20. Davoudi A, Wanigatunga AA, Kheirkhahan M, Corbett DB, Mendoza T, Battula M, et al.
424 Accuracy of Samsung Gear S Smartwatch for Activity Recognition: Validation Study. *JMIR*
425 *Mhealth Uhealth*. 2019;7(2):e11270.
- 426 21. John D, Miller R, Kozey-Keadle S, Caldwell G, Freedson P. Biomechanical examination
427 of the 'plateau phenomenon' in ActiGraph vertical activity counts. *Physiol Meas*.
428 2012;33(2):219-30.
- 429 22. Published cut-points and how to use them in GGIR: GGIR; [Available from:
430 <https://cran.r-project.org/web/packages/GGIR/vignettes/CutPoints.html>].
- 431 23. Esliger DW, Rowlands AV, Hurst TL, Catt M, Murray P, Eston RG. Validation of the
432 GENEActiv Accelerometer. *Med Sci Sports Exerc*. 2011;43(6):1085-93.
- 433 24. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients
434 for reliability research. *Journal of chiropractic medicine*. 2016;15(2):155-63.
- 435 25. Akoglu H. User's guide to correlation coefficients. *Turk J Emerg Med*. 2018;18(3):91-3.
- 436 26. Dancy CP, Reidy J. *Statistics without maths for psychology*: Pearson education; 2007.
- 437 27. Altman DG. *Practical statistics for medical research* Chapman and Hall. London and New
438 York. 1991.
- 439 28. POWELL SM, JONES DI, ROWLANDS AV. Technical Variability of the RT3
440 Accelerometer. *Medicine & Science in Sports & Exercise*. 2003;35(10):1773-8.
- 441 29. Nicolella DP, Torres-Ronda L, Saylor KJ, Schelling X. Validity and reliability of an
442 accelerometer-based player tracking device. *PLoS One*. 2018;13(2):e0191823.
- 443 30. Vanhelst J, Fardy PS, Beghin L. Technical variability of the Vivago® wrist-worn
444 accelerometer. *J Sports Sci*. 2014;32(19):1768-74.
- 445 31. : ActiGraph; [cited 2023 03/10/2023]. Available from:
446 <https://actigraphcorp.com/actigraph-wgt3x-bt/>.
- 447 32. Isakeit T. Fitbit Sense Teardown 2021 [Available from:
448 <https://www.ifixit.com/Teardown/Fitbit+Sense+Teardown/137130>].
- 449 33. Hildebrand M, Hansen BH, van Hees VT, Ekelund U. Evaluation of raw acceleration
450 sedentary thresholds in children and adults. *Scandinavian Journal of Medicine & Science in*
451 *Sports*. 2017;27(12):1814-23.
- 452 34. HILDEBRAND M, VAN HEES VT, HANSEN BH, EKELUND U. Age Group
453 Comparability of Raw Accelerometer Output from Wrist- and Hip-Worn Monitors. *Medicine &*
454 *Science in Sports & Exercise*. 2014;46(9):1816-24.
- 455 35. Kim Y, Crouter SE, Lee JM, Dixon PM, Gaesser GA, Welk GJ. Comparisons of
456 prediction equations for estimating energy expenditure in youth. *J Sci Med Sport*. 2016;19(1):35-
457 40.
- 458 36. Crouter SE, Churilla JR, Bassett DR, Jr. Estimating energy expenditure using
459 accelerometers. *Eur J Appl Physiol*. 2006;98(6):601-12.
- 460 37. Rowlands AV, Olds TS, Hillsdon M, Pulsford R, Hurst TL, Eston RG, et al. Assessing
461 sedentary behavior with the GENEActiv: introducing the sedentary sphere. *Med Sci Sports*
462 *Exerc*. 2014;46(6):1235-47.

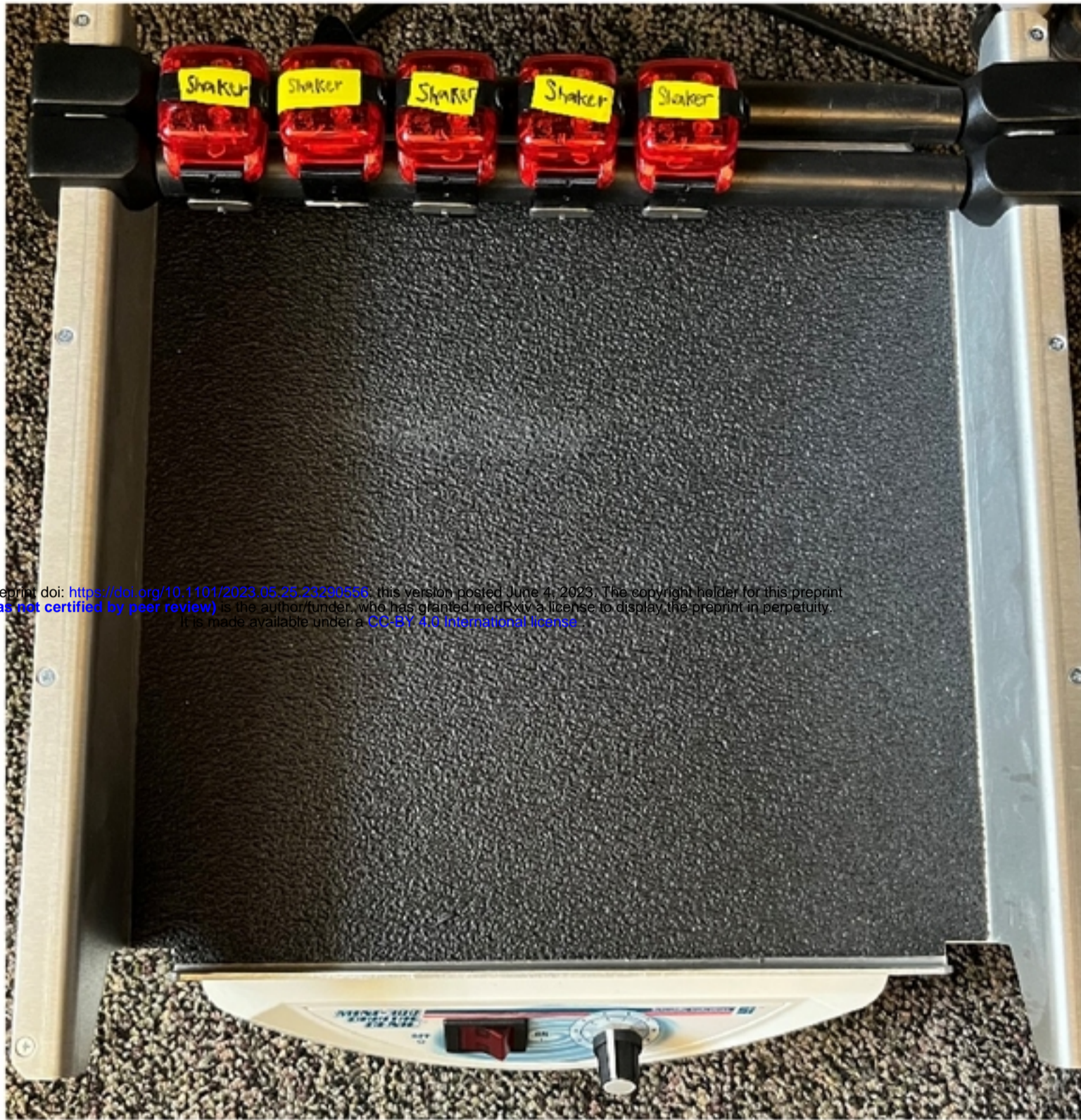
- 463 38. Shei RJ, Holder IG, Oumsang AS, Paris BA, Paris HL. Wearable activity trackers-
464 advanced technology or advanced marketing? Eur J Appl Physiol. 2022;122(9):1975-90.
465 39. Apple Watch models: Apple; [Available from: <https://www.apple.com/watch/compare/>].

466 **Supporting information**

467 **S1 Fig. Absolute Error of the Raw Acceleration Data from all Devices by Speed Compared**
468 **to the Accelerations Produced by a Mechanical Shaker Table.**

469 **S2 Fig. Mean Bias of the Raw Acceleration Data from all Devices by Speed Compared to**
470 **the Accelerations Produced by a Mechanical Shaker Table.**

471



medRxiv preprint doi: <https://doi.org/10.1101/2023.05.25.23291556>; this version posted June 4, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Fig1

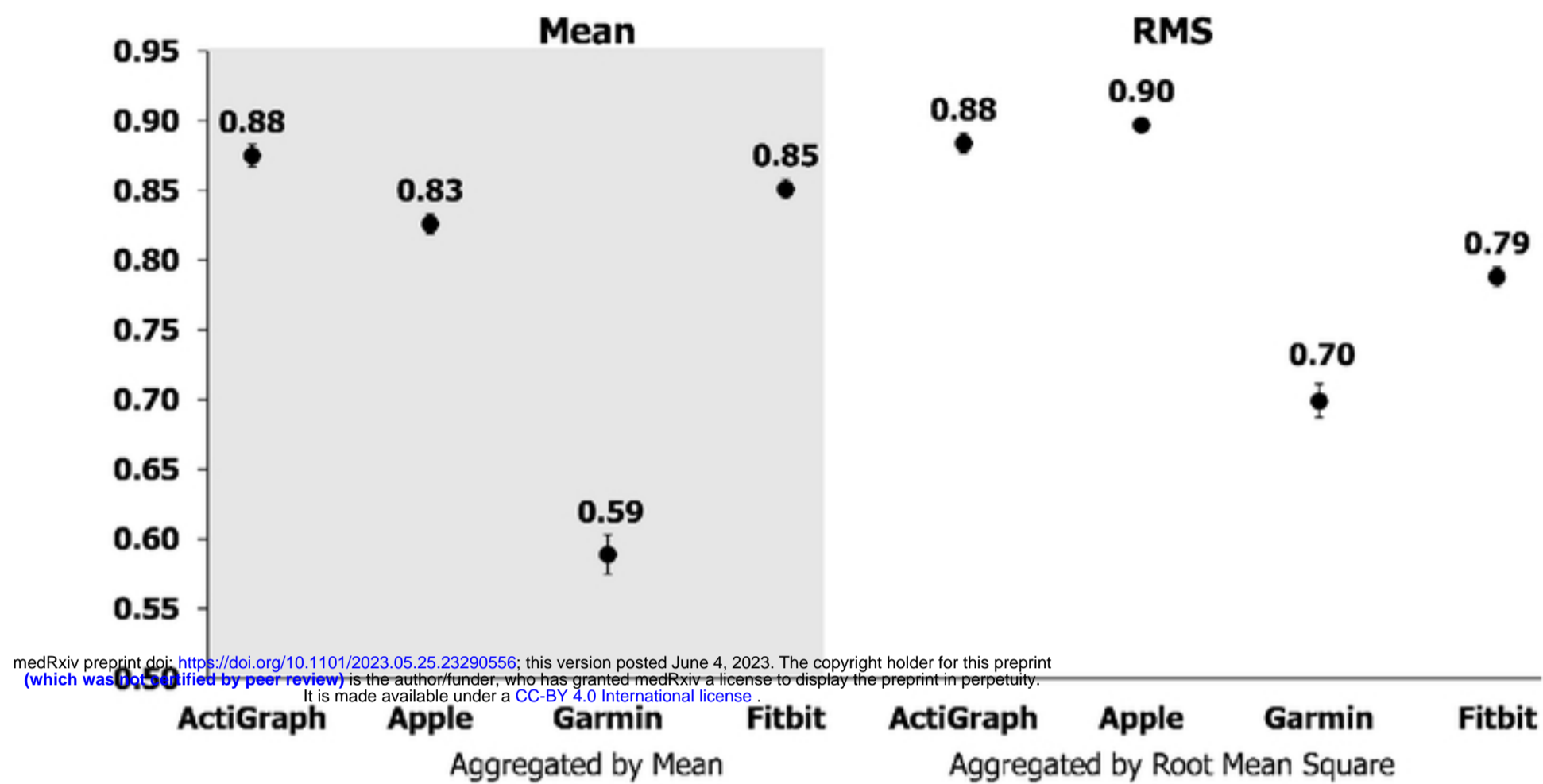
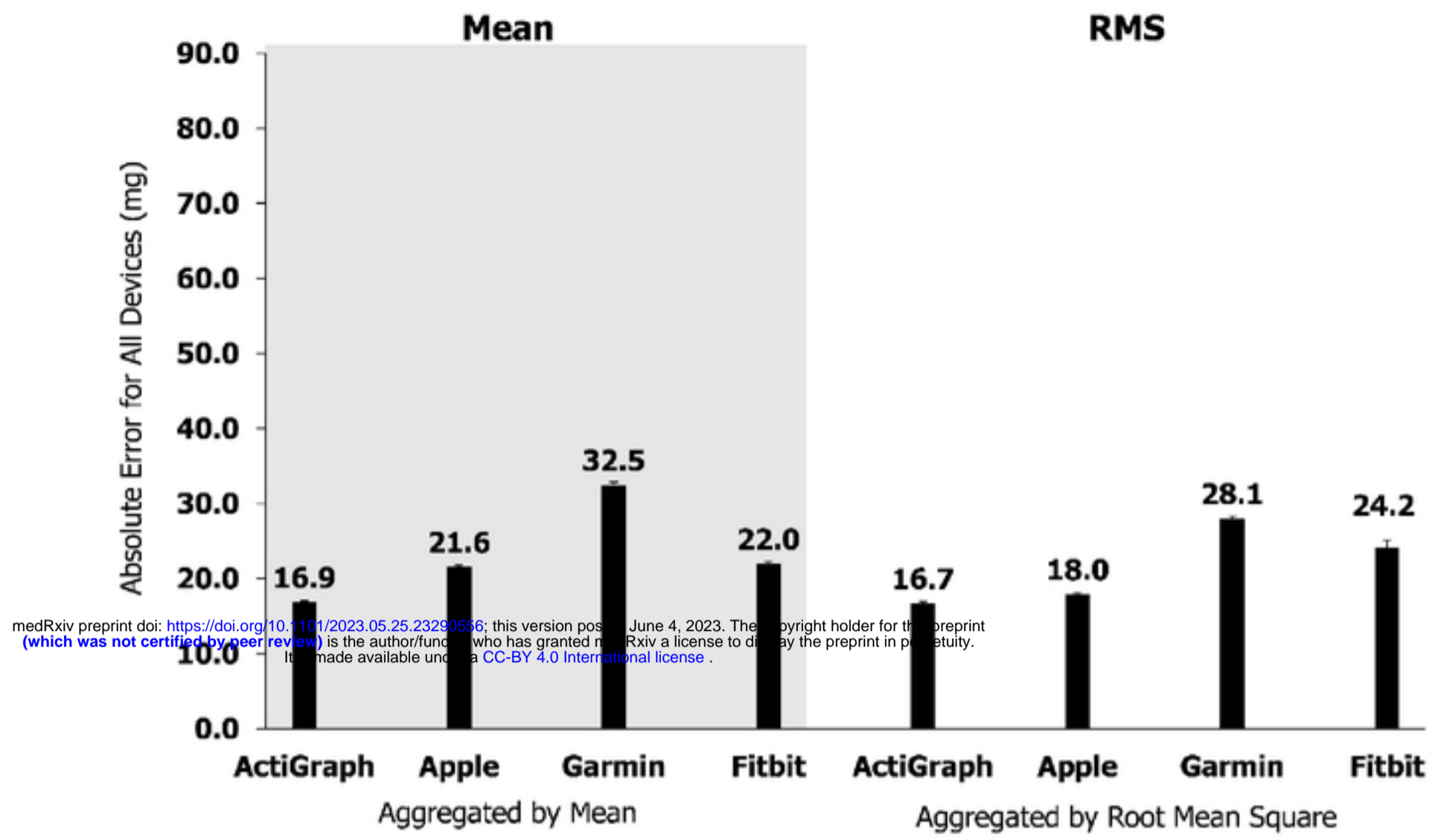
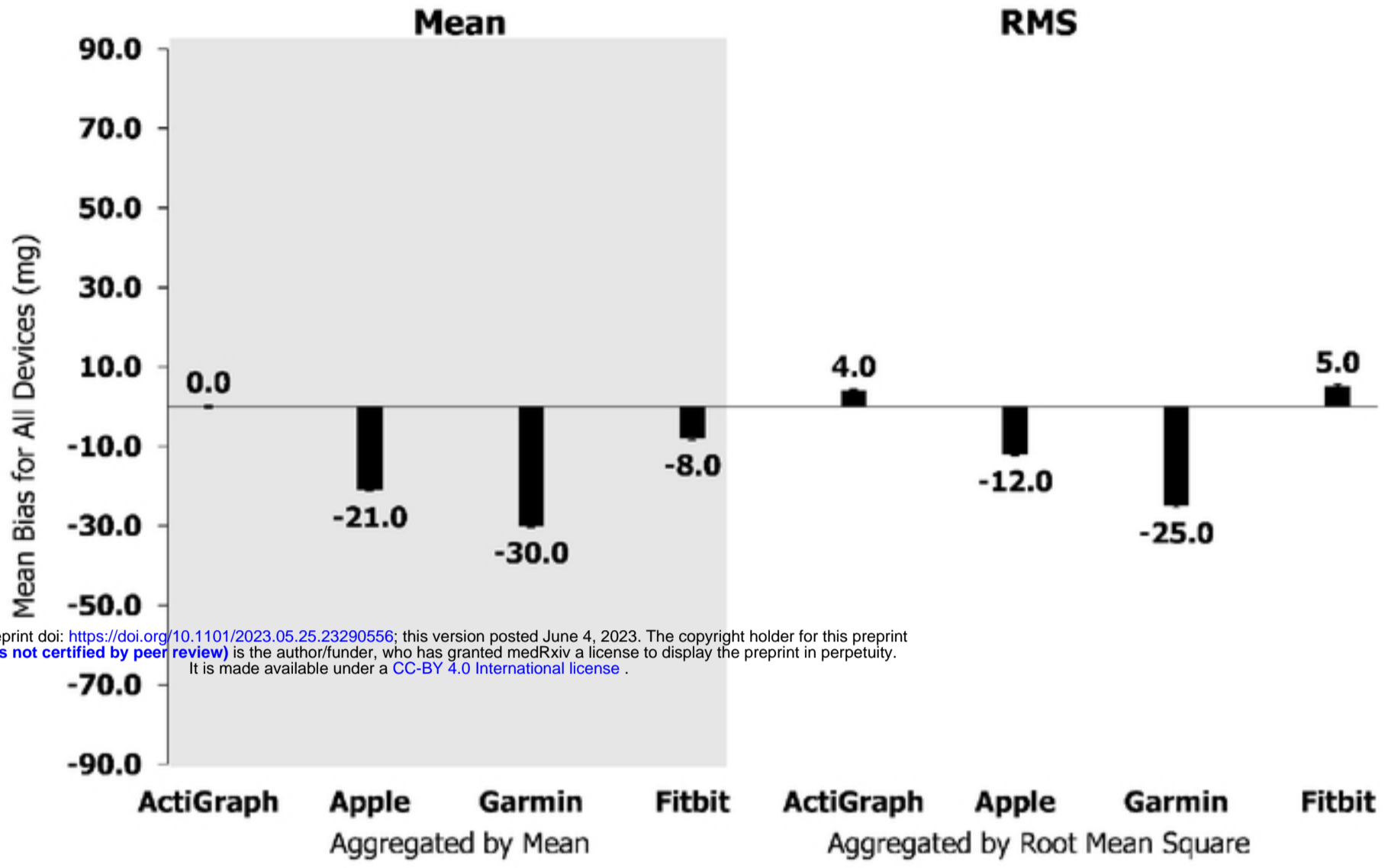


Fig2



medRxiv preprint doi: <https://doi.org/10.1101/2023.05.25.23290586>; this version posted June 4, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Fig3



medRxiv preprint doi: <https://doi.org/10.1101/2023.05.25.23290556>; this version posted June 4, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a [CC-BY 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Fig4