PLOS DIGITAL HEALTH

RESEARCH ARTICLE

Non-equivalent, but still valid: Establishing the construct validity of a consumer fitness tracker in persons with multiple sclerosis

Ashley Polhemus[1]*, Chloé Sieber[1,2], Christina Haag[1,2], Ramona Sylvester[3], Jan Kool[3], Roman Gonzenbach[3], Viktor von Wyl[1,2]

1 Epidemiology, Biostatistics, and Prevention Institute, University of Zurich, Zurich, Switzerland, **2** Institute for Implementation Science in Health, University of Zurich, Zurich, Switzerland, **3** Research Department Physiotherapy, Rehabilitation Centre, Valens, Switzerland

* ashley.polhemus@uzh.ch

Citation: Polhemus A, Sieber C, Haag C, Sylvester R, Kool J, Gonzenbach R, et al. (2023) Non-equivalent, but still valid: Establishing the construct validity of a consumer fitness tracker in persons with multiple sclerosis. PLOS Digit Health 2(1): e0000171. https://doi.org/10.1371/journal.pdig.0000171

Editor: Nicole Yee-Key Li-Jessen, McGill University, CANADA

Received: June 14, 2022

Accepted: November 23, 2022

Published: January 25, 2023

Data Availability Statement: Analytical data files are available on the Open Science Foundation project repository "BarKA: Fitbit validation study" (https://doi.org/10.17605/OSF.IO/V2B9E).

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Tools for monitoring daily physical activity (PA) are desired by persons with multiple sclerosis (MS). However, current research-grade options are not suitable for longitudinal, independent use due to their cost and user experience. Our objective was to assess the validity of step counts and PA intensity metrics derived from the Fitbit Inspire HR, a consumer-grade PA tracker, in 45 persons with MS (Median age: 46, IQR: 40–51) undergoing inpatient rehabilitation. The population had moderate mobility impairment (Median EDSS 4.0, Range 2.0–6.5). We assessed the validity of Fitbit-derived PA metrics (Step count, total time in PA, time in moderate to vigorous PA (MVPA)) during scripted tasks and free-living activity at three levels of data aggregation (minute, daily, and average PA). Criterion validity was assessed though agreement with manual counts and multiple methods for deriving PA metrics via the Actigraph GT3X. Convergent and known-groups validity were assessed via relationships with reference standards and related clinical measures. Fitbit-derived step count and time in PA, but not time in MVPA, exhibited excellent agreement with reference measures during scripted tasks. During free-living activity, step count and time in PA correlated moderately to strongly with reference measures, but agreement varied across metrics, data aggregation levels, and disease severity strata. Time in MVPA weakly agreed with reference measures. However, Fitbit-derived metrics were often as different from reference measures as reference measures were from each other. Fitbit-derived metrics consistently exhibited similar or stronger evidence of construct validity than reference standards. Fitbit-derived PA metrics are not equivalent to existing reference standards. However, they exhibit evidence of construct validity. Consumer-grade fitness trackers such as the Fitbit Inspire HR may therefore be suitable as a PA tracking tool for persons with mild or moderate MS.

Author summary

Physical activity (PA) is an important aspect of health and well-being. However, PA is often reduced in persons with multiple sclerosis (MS), a neurodegenerative autoimmune

disease which affects physical function, motor control, and energy levels. It is of public health interest to increase PA behavior in this population. However, valid and user-friendly methods for tracking PA are required to quantify PA behavior during patients' daily lives. So-called "research-grade" wearable devices are used for short-term measurements (for example, 7 days), but offer poor user experience and are therefore not suitable for longer-term PA tracking. It is therefore increasingly common for MS researchers to use "consumer-grade" devices such as Fitbits. However, high-quality evidence of their validity in MS populations is limited. In this study, we compared PA metrics derived from a Fitbit device to multiple, validated research-grade methods. While the PA metrics derived from each method were not equivalent, all exhibited the similar evidence of validity. In some cases, Fitbit outperformed research-grade methods. We posit that PA metrics derived from the Fitbit are now suitable for long-term PA tracking in MS populations, and that the resulting longitudinal data has the potential to progress our understanding of world PA behavior in MS populations.

Introduction

Multiple sclerosis (MS) is a neurodegenerative autoimmune disease which affects physical and cognitive function, motor control, and energy levels. Physical activity (PA) is often reduced in persons with MS (PwMS) [1,2], though it is known to aid in symptom and fatigue management [3–5] and is perceived as an important part of health care by PwMS [6,7]. Managing appropriate amounts of PA is often difficult for PwMS, as overexertion can cause severe short-term fatigue or symptom exacerbations before the benefits of PA are realized [8–10]. To enable the best health outcomes, tools for managing PA and fatigue are desired by PwMS [11].

For such tools to be effective, they must reliably and conveniently track PA over long periods of time, yielding either clinically or personally meaningful information. Consumer-grade PA trackers such as wrist-worn Fitbits are therefore gaining popularity in this population, and have already been used to generate PA outcomes in several large cohort and interventional studies [12–14]. They are easy to use, engaging, inexpensive, and provide meaningful PA metrics which are interpretable within the context of public health guidelines [15]. In addition, these devices enable users to interact with their own data, set goals, and review progress over time. These features promote long-term engagement with remote monitoring technologies [16,17]. The resulting rich, longitudinal data could provide insights into PA behavior not observed in traditional periodic or questionnaire-based PA metrics.

However, only limited evidence of validity is available for any Fitbit device in MS populations. Existing validation studies are primarily conducted in healthy adults, and three recent systematic reviews of such studies cautiously support the validity of Fitbit-derived PA metrics [18–21]. However, validation studies also suggest that these metrics' accuracies decrease at low activity intensities [20], at slow walking speeds [18,22–24], and with the use of walking aids [25]. Not only do PwMS walk slower healthy controls, they also exhibit different abnormal gait patterns [26,27] and frequently adopt walking aids as their MS progresses [28,29]. It is plausible that these factors affect the validity of Fitbit-derived PA metrics in PwMS. To date, validation studies in PwMS are limited to step count, and do not address the other PA metrics produced by these trackers [30,31]. Given the expanding use of wrist-worn Fitbits to track PA in MS, a thorough evaluation of their validity in this population is warranted.

In this study, we aimed to expand and update existing evidence on the validity of wrist-worn Fitbit devices in MS populations. We assessed the construct validity of three PA metrics–step count, time spent in PA, and time spent in moderate to vigorous PA (MVPA)–derived

from the Fitbit Inspire HR. We did this by comparing Fitbit-derived PA metrics to multiple reference measures (Table 1), and systematically triangulating evidence of their criterion validity, convergent validity, and known-groups validity (Fig 1). This validation study evaluated PA metrics according to validation best practices, accounting for the known shortcomings of existing reference measures [32].

Materials and methods

Objective

The objective of this study was to assess the construct validity of physical activity (PA) metrics derived from the Fitbit Inspire HR, a consumer-grade fitness tracker. Construct validity is the extent to which an index measure–or the instrument under study–measures the theoretical construct it is supposed to measure [46]. Several sub-types of validity comprise construct validity [47]. In this study, we assess Fitbit-derived PA measures in terms of their criterion validity, known-groups validity, and convergent validity. Criterion validity refers to an instrument's ability to measure the concept it purports to measure, and is typically assessed through correlations and agreement with a well-validated reference standard, or "criterion measure." [48] Known-groups validity is the ability of an instrument to discriminate between groups of individuals which are known to differ from each other, such as disease severity strata [49]. Finally, convergent validity refers to a measure's ability to demonstrate an expected relationship with other theoretically related, clinically relevant constructs [50]. Convergent validity is often assessed through correlation and other association measures.

This validation study was conducted as part of BarKA-MS, a cohort study on the barriers and facilitators to PA in PwMS [51]. It expands upon best practices developed by Johnston et al., [32] who propose a six-step framework for designing and reporting validation studies of consumer wearables: 1) target population, 2) index measure (the measure being validated), 3) testing conditions, 4) criterion measure (the reference standard), 5) data processing methods, and 6) statistical analysis.

Target population

Our target population was ambulatory PwMS. We recruited a convenience sample of PwMS undergoing elective inpatient neurorehabilitation at the Kliniken Valens between January and November 2021. Participants were eligible if they 1) had a confirmed diagnosis of MS according to their referring physician, 2) were 18 years of age or older, 3) had reduced walking ability but were able to walk independently with or without an assistive device, 4) had access to WiFi and a mobile device in the rehabilitation center and at home, 5) were willing to wear study devices to measure their PA, and 6) were able to answer study questionnaires in German. The BarKA-MS study was composed of two phases (in the clinic and at home). The first phase lasted between one to three weeks depending on the length of the rehabilitation stay and the second phase lasted four weeks. We set a target sample size of 45 participants based on the expected rate of enrollment at Kliniken Valens in the first half of 2021. The recruitment window was then extended due to slower than expected enrollment throughout the COVID pandemic. The ethics committee of the canton of Zurich approved the study protocol (BASEC-no. 2020–02350) and all participants provided written informed consent.

Index measure

Our index measures–or the measures we aimed to validate–were step count, time in PA, and time in MVPA derived from the Fitbit Inspire HR. The Fitbit Inspire HR is a consumer PA

Table 1. Methods used to triangulate the validity of Fitbit-derived PA metrics.

Method	Description
Step count	
Manual	Scripted tasks were video-recorded and two assessors manually counted steps according to a validated standard operating procedure. The two assessors' counts were averaged. Manual counts were used as criterion measures during scripted tasks only.
Actigraph (Standard)	During post-processing, a band pass filter is applied to Actigraph's raw accelerometer signal to remove movement artifact outside the range of human motion. Actigraph's step count algorithm detects footfalls by identifying peaks in the accelerometer signal, and can therefore be affected by choice of filter. The standard filter was developed by the manufacturer for use in healthy populations, but has also been used in populations with MS. It is known to underestimate step count, especially in populations with walking impairments [33,34].
Actigraph (LFE)	The Low Frequency Extension (LFE) is a revised band pass filter which enhances the Actigraph's sensitivity to slow movements. It is recommended by the manufacturer in populations with impaired mobility, and is also frequently used in MS populations. The LFE has been shown to increase sensitivity to slow stepping in laboratory settings. However, it is also known to overestimate step count during free-living activity [33].
Fitbit	Fitbit's proprietary step detection algorithm derived step count from the device's raw accelerometer signal. Fitbit provides step counts at up to minute-level granularity through its application programming interface (API).
Time in PA	
Actigraph (Vertical)	PA intensity is derived from the Actigraph by applying cutpoints to the number of activity counts identified per minute. The Actigraph (Vertical) method differentiates between sedentary behavior and PA depending on whether a minimum cutpoint of 100 vertical axis counts per minute is met [35]. This method was validated on an older model of Actigraph which is roughly equivalent to data derived from the LFE on the Actigraph GT3X. It is widely used in MS populations [36,37].
Actigraph (VM)	The Actigraph Vector Magnitude (VM) method accounts for three-dimensional motion, rather than motion in the vertical axis. This method categorizes minutes as sedentary or PA with a cutpoint of 150 VM counts per minute. It is typically used with the standard filter applied. It was derived from healthy individuals [35,38,39], but is also used in persons with MS [40,41].
Fitbit	Fitbit's proprietary PA classification algorithm classifies minutes into four intensity categories: sedentary, lightly active, fairly/moderately active, and very active [42]. These categories loosely align with those used by research-grade devices: sedentary, light PA, moderate PA, vigorous PA. We defined all non-sedentary minutes as time in PA. The factors which influence PA classification are not publicly available. However, movement intensity, heart rate, and breathing rate are mentioned by moderators of Fitbit's support forum [43]. Fitbit provides PA classifications at the minute level through its API, which may then be aggregated into hourly or daily metrics.
Time in MVPA	
Actigraph (Uniform)	The Actigraph (Uniform) method was developed to differentiate between light and MVPA in populations with MS and was validated during over-ground walking at multiple speeds. It defines minutes which exceed 1745 vertical axis counts to be MVPA, and uses the LFE filter [44].
Actigraph (Severity)	The Actigraph (Severity) method was developed alongside the Actigraph (Uniform) method, but proposes different cutpoints according to MS severity: • Mild/moderate MS (EDSS < 6.0): 1980 vertical counts per minute • Severe MS (EDSS > = 6.0): 1185 vertical counts per minute Both cutpoints are designed to be used with the LFE filter [44].
Actigraph (Sasaki)	The Actigraph (Sasaki) method uses a cutpoint of 2690 VM counts per minute to differentiate between light and MVPA. This method was developed in healthy controls during treadmill walking at varying speeds [38]. It has since been used in populations with MS [40,45].
Fitbit	We defined all minutes characterized as 'fairly/moderately active' and 'very active' by Fitbit's proprietary algorithm as time in MVPA [42]. This grouping aligns with the output of the Actigraph Uniform, Severity, and Sasaki methods, which do not differentiate between moderate PA and vigorous PA [38,44].

MS: multiple sclerosis; PA: physical activity; LFE: Low frequency extension; MVPA: moderate to vigorous physical activity; EDSS: expanded disability status score

https://doi.org/10.1371/journal.pdig.0000171.t001

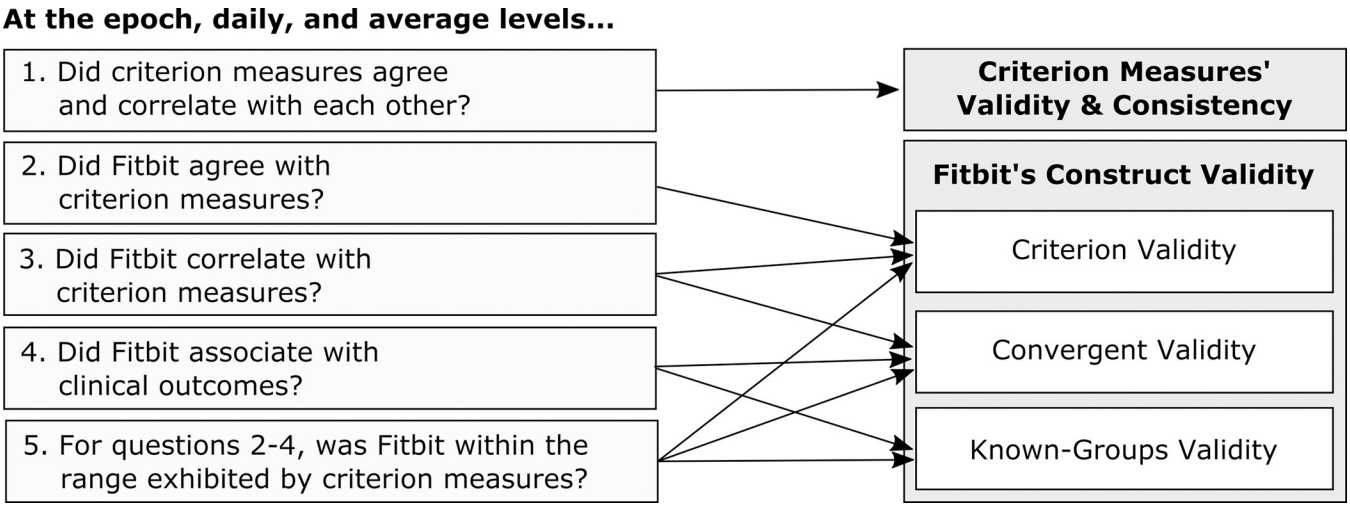


Fig 1. Research questions addressed in this validation study.

https://doi.org/10.1371/journal.pdig.0000171.g001

tracker which is worn on the wrist and measures step count, PA intensity, sleep, heart rate, and other fitness metrics at up to minute-level granularity. Participants were given a Fitbit Inspire HR and were instructed to wear it on their non-dominant wrist during the day and if desired at night throughout the course of the study. The accompanying mobile application was installed on each participant's mobile device, and each participant was given a de-identified, pre-configured study account. Alerts and daily goals were either turned off or set to a minimum, and participants were encouraged to leave these settings off for the duration of the study. Minute-level data were collected and stored through the Fitabase platform (Fitabase, San Diego, California), a cloud-based study management platform which provides industry-standard security measures such as encryption, password protection, access logs, etc. All participants consented to the privacy statements and settings associated with these platforms.

Testing conditions

According to Johnston et al.'s framework, index measures were compared to criterion measures during laboratory evaluation (i.e., controlled walking tests), semi-free-living evaluation (i.e., scripted assessments which simulate various free-living activities), and free-living evaluation (i.e., during daily living 'in the wild') [32]. For brevity, we refer to laboratory evaluations and semi-free living evaluations together as 'scripted tasks.'

Laboratory evaluation

Rehabilitation schedule permitting, PA metrics were assessed manually, via the Fitbit, and via criterion measures during a 6-Minute Walk Test [52] in participants' final week at the clinic. Criterion measures are described in greater detail in the next section. All participants were instructed to cover as much distance in six minutes as possible, and rests were allowed. Participants rested in a seated position for at least three minutes immediately prior to and following the test to allow for confirmation of timestamp alignment between devices.

Semi-free-living evaluation. A sub-sample of participants also completed an assessment comprised of five scripted tasks designed to replicate movement patterns regularly encountered in daily life. PA metrics were assessed via the Fitbit and via criterion measures (see below) during these tasks. The semi-free-living evaluation consisted of:

- *Walking with postural transitions*: Participants repeatedly rose from a seated position, walked approximately five meters to an examination bed, lay supine for three seconds, returned to the chair, and sat for three seconds. This task was designed to assess the effect of short walking bouts interrupted with postural transitions.
- *Simulated cleaning*: Participants repeatedly carried a series of glasses, cups, saucers, and towels from one table to another nearby table. During each repetition, participants unfolded and re-folded the towels. This task simulated light PA with short walking bouts in a confined space, frequent direction changes, and weight shifting between legs. We designed this task to simulate working in a kitchen or tidying a room.
- *Sit to stand*: In this task, participants repeatedly rose from and returned to a seated position. This activity further tested how postural transitions are characterized by index and criterion measures.
- *Wheelchair push*: Participants propelled themselves around a circular track in a wheelchair with the Fitbit worn on the outermost wrist to assess how manual wheelchair propulsion, and more generally upper body activity, is characterized.
- *Stair climb and descent*: In this task, participants repeatedly walked up and down two flights of stairs to assess step count accuracy during stair climbing and descent.

These activities were selected and designed in collaboration with subject matter experts at the rehabilitation facility. Each semi-free-living evaluation lasted approximately 30 minutes. Participants were instructed to complete each task at a pace they could maintain safely for three minutes and to use their preferred walking aids. Rests were allowed. Participants rested in a seated position for at least three minutes immediately prior to and following each task to enable confirmation of timestamp alignment and to mitigate fatigue effects.

Free-living evaluation. For the purposes of this evaluation, participants wore both the Fitbit and a criterion measure (Actigraph GT3X, see below) under free-living conditions for approximately 14 days. This two-week period was comprised of their final week in the rehabilitation clinic and the following week in their home environment. Participants occasionally wore the devices longer if the rehabilitation period was unexpectedly extended. After participants had worn the device at home for seven days, the participants logged the dates they had worn the devices and returned the Actigraph GT3X to investigators by mail. Participants continued to wear the Fitbit as part of the BarKA cohort study.

Criterion measures

Average manual step counts were considered the criterion measure for assessing Fitbit's step count algorithm during scripted tasks. Tasks were video-recorded and two assessors manually counted steps according to a validated standard operating procedure (S1 Text).

Several additional criterion measures were derived from the Actigraph GT3X (Manufacturing Technology, Inc., FL, USA), a research-grade accelerometer which has been validated in PwMS [53,54]. Actigraph devices were initialized in Actilife 6.0 with a sampling rate of 30Hz and worn on the right hip. However, multiple data processing methods exist to derive PA metrics in this population (Table 1) [38,44,55]. These methods use different data (i.e., 1-dimensional vs. 3-dimensional movement) and processing methods (i.e., standard vs. highly sensitive filtering) to calculate PA metrics. However, the Fitbit is not expected to agree exactly with any of the criterion measures derived from the Actigraph GT3X (Table 1). The Actigraph measures were derived and validated for wear on the hip [35,38,44], whereas Fitbit is wrist-worn. The Actigraph GT3X-based methods derive PA metrics from an accelerometer only

[35,38,44]. The factors which influence PA classification are not publicly available, though support resources suggest that movement intensity, heart rate, and breathing rate may influence PA estimation [43]. Finally, Actigraph-derived measures are non-equivalent with each other [56]. Any Actigraph method may therefore impart criterion standard bias if compared to Fitbit as a single criterion measure [57].

We therefore opted to assess the metrics derived from Fitbit through triangulation [58] in an agreement validation study [57] and through an assessment of construct validity. Criterion measures for step count, time in PA, and time in MVPA were derived from Actigraph through multiple established methods (Table 1). Two Actigraph-based methods were used to derive step count (referred to as Actigraph (Standard) and Actigraph (LFE)) [59], two methods were used to derive time in PA (Actigraph (Vert) [35] and Actigraph (VM)) [38], and three methods were used to derive time in MVPA (Actigraph (Uniform), [44] Actigraph (Severity), [44] and Actigraph (Sasaki)) [38].

Construct validity was further evaluated by quantifying the relationship between PA metrics and theoretically-related clinical assessments. Convergent validity was assessed through associations with patient reported outcomes and clinical outcome measures. Patient reported outcomes included the MS Walking Scale– 12 (MSWS-12), a patient-reported measure of walking ability and its impact on daily activities [60,61] and the International PA Questionnaire (IPAQ), a self-assessment of PA during the previous seven days [62]. Clinical measures included the Expanded Disability Status Scale (EDSS) [63]; the 10-meter Gait Speed test (10mGS) [64]; and the 6-Minute Walk Test (6MWT) [65]. These measures were assessed during the last week of rehab, except for the IPAQ, which was reported by participants following the free-living assessment. Known-groups validity was assessed by comparing PA metrics between subgroups according to established cutoffs of clinical scales. Disease severity strata were defined as mild (EDSS < 4.0), moderate (EDSS 4.0–5.5), and severe (EDSS 6.0–6.5) body function impairment, consistent with previous studies [44].

Data processing

Actigraph data were uploaded to Actilife, filtered to remove non-human movement artifact with both the standard filter and the low frequency extension (LFE), aggregated into one minute epochs, and exported for further processing. Step count, PA intensity (sedentary behavior, LPA, MVPA), and heart rate data derived from the Fitbit Inspire HR were calculated according to Fitbit's proprietary algorithms and extracted in one minute epochs. All processing was conducted in R, version 4.1.0 in the RStudio environment, version 1.4.1717. Validated algorithms (Table 1) were applied to derive PA intensity and step count.

Non-wear time was defined as 30 minutes of continuous zeros with a 2-minute spike tolerance [66]. For Actigraph, this definition referred to epochs with zeroes in the x, y and z axes, and for Fitbit this referred to epochs with zero step count, sedentary PA categorization, and no registered heart rate. Wear periods shorter than 10 minutes were discarded to reduce false positives in wear time resulting from short spikes. Days with at least 10 hours of wear time during waking hours were considered valid [67], and participants with at least two valid days were included in this analysis [68]. Epochs in which both devices were worn during waking hours (6AM to 11PM) on valid days were included in aggregation and analysis. Data categorized as non-wear time and epochs which occurred on non-valid days were removed. The day participants left the clinic and traveled home was not included in this analysis, as these days did not represent 'normal' activity. To limit the effects of differential wear patterns on agreement analyses, only minutes during which both the Fitbit and the Actigraph were worn were included in data aggregation and further analysis.

Data aggregation. For each method, PA data were then aggregated into three levels of granularity for agreement analysis: "epoch-level", "daily", and "average" PA. Epoch-level data was used to evaluate absolute agreement between PA metrics over short periods of time and during diverse activities of daily living. Timestamp alignment within one minute was confirmed according to visit notes, videos, and manual inspection for each participant. Minute-level step counts were aggregated into 5 minute epochs. An agreement window of plus or minus one minute was applied in a pairwise fashion to minute-level PA intensity metrics. This window accounted for the effects of timestamp misalignment and the potential dependency of Fitbit's PA algorithm on heart rate. An epoch was considered in agreement if Fitbit-derived PA intensity yielded the same categorization as Actigraph-derived PA intensity within a window of plus or minus one minute of the Actigraph's timestamp. Daily PA metrics were calculated by summing all included minute-level data per patient per day. Days in both the rehab setting and the home setting were included in analyses at the daily level of aggregation. Average-level PA metrics were calculated for the home environment only by averaging each participant's daily PA metrics over all valid days, consistent with previous PA study outcomes in MS populations [40,69].

Data labeling. Data collected during laboratory and semi-free-living evaluations were extracted and labeled by consulting visit notes, video timestamps, and manual inspection. Manual and device-derived step counts were calculated for each scripted task and for the rests between tasks.

Statistical analysis

Agreement of categorical data was assessed through a multi-level implementation of Fleiss' kappa assuming participant-level random effects [70]. Differences in PA categorizations during individual scripted tasks were identified through Fisher exact tests. Kruskal-Wallis tests, Wilcoxon signed-rank tests, Pearson product-moment correlation coefficients (Pearson's r), Lin's concordance correlation coefficients (CCC) [71] evaluated the differences, correlations, and absolute agreement between measures for continuous and count data. Bland Altman plots [72] visualized the mean bias and limits of agreement at the daily level. At the epoch and daily level, Pearson's r, CCC, and Bland Altman statistics were adjusted for patient-level random effects according to the procedures defined by Parker et al. [73] Pearson's r was selected because it was adjustable for patient-level random effects, and data were visually assessed for approximate normal distributions. Confidence intervals were derived through bootstrapping. In sensitivity analysis, these analyses were repeated for each disease severity stratum. For data collected during scripted tasks, this analysis was conducted for all scripted tasks together, accounting for task-level random effects as described by Parker et al. [73] Wilcoxon-Mann-Whitney tests and Wilcoxon effect sizes [74] quantified the existence and magnitude of differences across known groups. Pearson's r quantified the relationships between average PA metrics and clinical measures.

Triangulation

"Triangulation" refers to the use of more than one approach to address a research question. By combining multiple methods and comparing results from different perspectives, the limitations of each method individually can be contextualized and addressed [58]. We developed a qualitative triangulation process to assess Fitbit's construct validity relative to several criterion measures at several levels of data aggregation. We did this by qualitatively considering the devices' performance during each evaluation according to a pre-defined list of considerations (Fig 1).

Each of these research questions was addressed at five levels of data aggregation: during scripted tasks, at the epoch, daily, and average level, and across the three disease severity strata. Analyses were rated according to the level of agreement and correlation exhibited by the various PA metrics. Categories aligned with widely accepted, though arbitrary, interpretations of correlations and kappa statistics found in the literature [75,76]:

- ++: Excellent agreement or strong correlations (0.75–1.0)
- +: Fair to good agreement or moderate correlations (0.4–0.74)
- -: Poor agreement, weak correlations (0.2–0.39)
- –: Very weak or complete lack of agreement or correlation (<0.2)
- ++/+/-/–: Evidence was mixed

Results

Participant characteristics

Of the 47 participants originally enrolled, two participants left rehabilitation early and had to be excluded from the study. Of the 45 remaining participants, 29 (64.4%) were female and 19 (35.6%) were male. The median age was 46 (IQR: 40–51) years. Median EDSS was 4.0 (Range: 2.0–6.5), indicating moderate disease severity of the population. Most participants had either secondary-progressive MS (42.2%) or relapsing-remitting MS (40%). The median time from diagnosis was 11 years (IQR 5–21). The participants who completed the 6MWT varied in average walking cadence (Median (Range): 109 (61–146) steps per minute) and walking aid use (none: 23, walking sticks: 10, rollator: 2). Due to rehabilitation schedules, Actigraph wear compliance, and Actigraph data corruption, not all participants were included in all evaluations. The number of participants included in each analysis are described in Fig 2, and their characteristics are described in S1 Table.

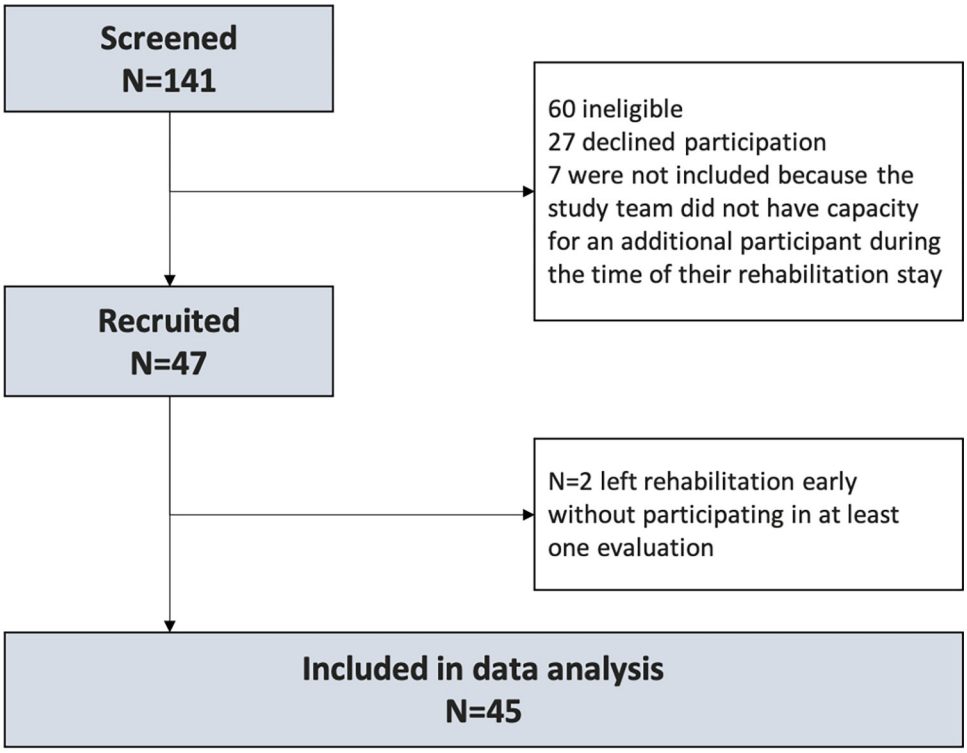
During free-living evaluations, participants wore Fitbits for an average of 16.4 (Standard deviation: 0.9) hours on 12.9 (1.9) valid days, whereas they wore the Actigraph for an average of 12.1 (0.9) hours on 8.6 (3.2) valid days.

Step count

1. Step count: Did criterion measures agree and correlate with each other?. During scripted tasks, both Actigraph methods–Actigraph (Standard) and Actigraph (LFE)–demonstrated strong correlation and good agreement with manual counts (r: 0.97–0.98; CCC: 0.68–0.73) and with each other (r: 0.97; CCC: 0.60) (Table 2). Actigraph (Standard) often underestimated step count compared to manual counts, and Actigraph (LFE) often mischaracterized non-walking movement and postural transitions as steps (Fig 3 and Fig 4).

During free-living evaluation, the two methods exhibited weak correlation and no agreement at the epoch level (r: 0.27; CCC: 0.03) and strong correlation but poor agreement at the daily (r: 0.88; CCC: 0.15) and average levels (r: 0.89; CCC: 0.27) (Table 2). These patterns were consistent across disease severity strata. (S2 Table)

2. Step count: Did Fitbit agree with criterion measures?. During scripted tasks, Fitbit-derived step counts demonstrated good agreement with manual counts (CCC: 0.66) and with Actigraph-derived counts (CCC (Standard): 0.55, CCC(LFE): 0.65) (Table 2). During scripted walking tasks, Fitbit-derived step counts were consistent with manual and Actigraph-derived counts, with one exception: Fitbit registered zero steps for two participants who used rollators during the test. Walking stick use did not appear to effect step count during the 6MWT or



Evaluation	N	Exclusion: N & reasons
Laboratory evaluation	35	8 due to scheduling constraints 1 Actigraph data file corrupted 1 Actigraph did not record
Semi-free-living evaluation	12	Target sub-group of 10-12 participants 13 recruited 1 Actigraph data file corrupted
Free-living evaluation (epoch-level aggregation)	42	2 did not meet minimum wear time requirements for Actigraph 1 Actigraph data file corrupted
Free-living evaluation (daily-level aggregation)	42	2 did not meet minimum wear time requirements for Actigraph 1 Actigraph data file corrupted
Free-living evaluation (average-level aggregation)	35	9 did not meet minimum Actigraph wear time requirements at home 1 Actigraph data file corrupted

Fig 2. Flowchart of recruitment and participation in this study.

https://doi.org/10.1371/journal.pdig.0000171.g002

other walking tasks. Upper-body movement was often mischaracterized as steps by Fitbit (Fig 4).

During free-living evaluation, agreement between Fitbit and Actigraph was substantially reduced compared to scripted tasks. However, step counts derived from the Fitbit consistently exhibited equivalent or higher agreement with each of the Actigraph methods than the Actigraph methods did with each other (Table 2). Fitbit produced significantly higher step counts

Table 2. Correlation and agreement between step counts derived from Fitbit and Actigraph.

Comparison	r [95% CI]	CCC [95% CI]
Scripted tasks		
Manual vs Act(Stand)[a]	0.97 [0.77–0.99]	0.68 [0.37–0.82]
Manual vs Act(LFE)[a]	0.98 [0.78–0.99]	0.73 [0.13–0.84]
Act(Stand) vs Act(LFE)[a]	0.97 [0.85–0.99]	0.60 [0.06–0.76]
Fitbit vs. Manual	0.92 [0.63–0.98]	0.66 [0.14–0.80]
Fitbit vs. Act(Stand)	0.92 [0.65–0.97]	0.55 [0.08–0.72]
Fitbit vs. Act(LFE)	0.93 [0.65–0.98]	0.65 [0.24–0.77]
Free living, Epoch level		
Act(Stand) vs Act(LFE) [a]	0.27 [0.24–0.30]	0.03 [0.02–0.04]
Fitbit vs. Act(Stand)	0.22 [0.19–0.25]	0.04 [0.03–0.05]
Fitbit vs. Act(LFE)	0.22 [0.19–0.25]	0.03 [0.02–0.05]
Free living, Daily level		
Act(Stand) vs Act(LFE) [a]	0.88 [0.84–0.91]	0.15 [0.11–0.20]
Fitbit vs. Act(LFE)	0.82 [0.78–0.87]	0.33 [0.22–0.43]
Fitbit vs. Act(Stand)	0.80 [0.75–0.85]	0.44 [0.32–0.57]
Free living, Average level		
Act(Stand) vs Act(LFE) [a]	0.89 [0.78–0.94]	0.27 [0.16–0.37]
Fitbit vs. Act(LFE)	0.86 [0.74–0.93]	0.50 [0.34–0.63]
Fitbit vs. Act(Stand)	0.82 [0.67–0.90]	0.65 [0.47–0.77]

[a] Comparison between two criterion measures

Act: Actigraph; Stand: Standard; LFE: Low frequency extension; r: Pearson correlation coefficient; CI: confidence interval; CCC: Lin's Concordance correlation coefficient

https://doi.org/10.1371/journal.pdig.0000171.t002

than Actigraph (Standard) and significantly lower counts than Actigraph (LFE) (both $p < 0.001$) (Fig 5A). At the epoch level, Fitbit exhibited no agreement with either Actigraph method (Table 2). Between two and four percent of epochs exhibited substantial inconsistency, with minimal steps detected by one device and large steps counts detected by the other (S1 Fig). Fitbit demonstrated poor to good agreement with each Actigraph method at the daily (CCC: 0.33–0.44) and average levels (CCC: 0.50–0.65) (Table 2). Bland Altman analysis yielded wide limits of agreement, further confirming the weak agreement of daily step counts (Fig 5B–5D). Agreement tended to be highest for participants with moderate MS and lowest for participants with severe MS (S2 Table).

3. Step count: Did Fitbit correlate with criterion measures?. During scripted tasks, Fitbit-derived step counts were strongly correlated with those derived from criterion measures (r (Standard): 0.92, r(LFE): 0.93) (Table 2).

During free-living evaluation, Fitbit and Actigraph-derived step count metrics were weakly correlated at the epoch level (r: 0.22), but strongly correlated at the daily (r: 0.80–0.82) and average levels (r: 0.76–0.86) (Table 2). At all levels of aggregation, correlations between Fitbit and Actigraph-derived step counts were similar in magnitude to that of the two Actigraph methods (Table 2). Trends were consistent across disease severity strata (S2 Table).

4. Step count: Did Fitbit associate with clinical outcomes?. Average step counts derived from the Fitbit during free-living evaluation exhibited moderate to strong correlations with most clinical measures (Table 3). These associations were similar in magnitude to those exhibited by Actigraph. Fitbit-derived step count also demonstrated the expected differences between subgroups, though effect sizes for Fitbit were lower than Actigraph-derived counts (Table 3).

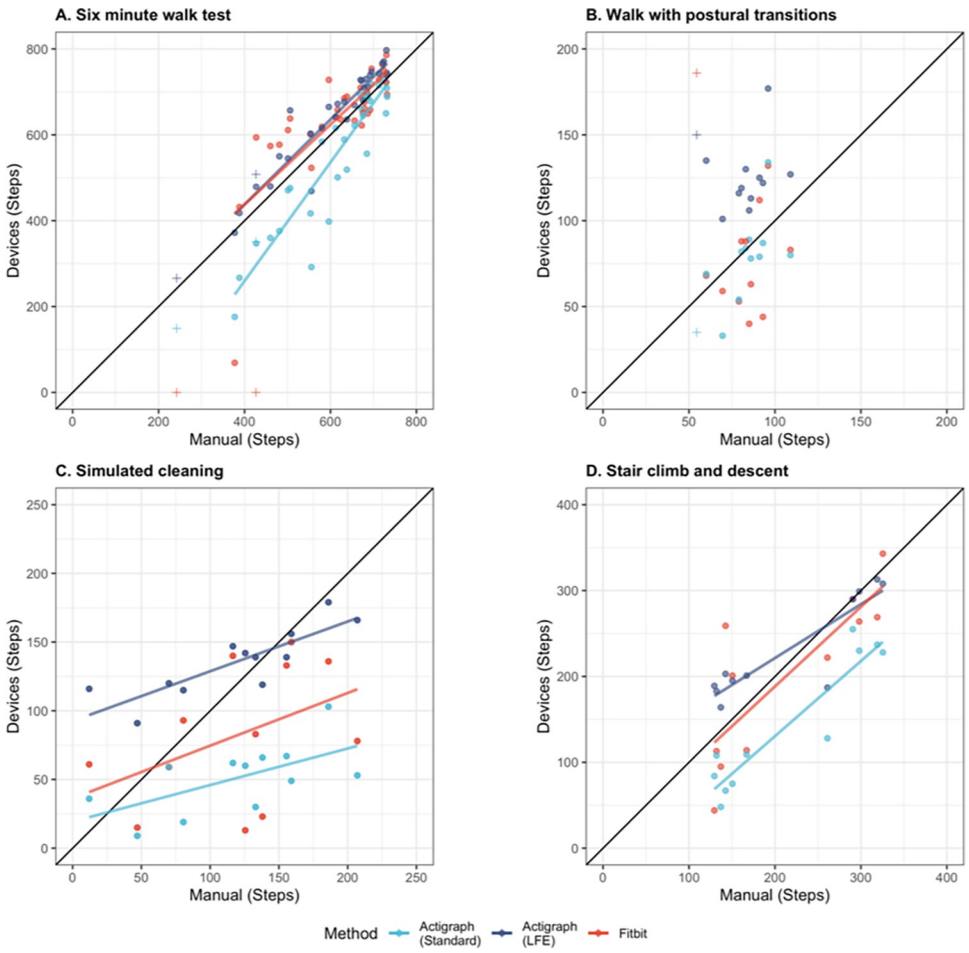


Fig 3. Manual and device-derived step counts during scripted walking tasks. Walking was assessed A) during a six-minute walk test, B) during walking interspersed with postural transitions, C) during a simulated cleaning task, and D) during stair climbing and descent. Measurements in perfect agreement would fall along the indicated diagonal black line. Regression lines (A, C, D) demonstrate deviations between actual and perfect agreement. Regression lines were omitted from B because the population did not exhibit sufficient variation in step count to yield linear trends. Fitbit registered no steps for two patients who used rollators (noted with +) during the 6-minute walk test (A). One participant's dyskinesia (noted with +) caused Fitbit to overestimate step count while walking with postural transitions (B), though not on other tasks when she could use a rollator or hold on to nearby structures for balance.

https://doi.org/10.1371/journal.pdig.0000171.g003

Time in physical activity

1. Time in PA: Did criterion measures agree and correlate with each other?. During scripted tasks, the two Actigraph methods–Actigraph (Vertical) and Actigraph (VM)–exhibited excellent epoch-level agreement (Fleiss' k: 0.93) (Table 4). During free-living evaluation, epoch-level agreement decreased slightly, but remained high (k: 0.75). At the daily level, the two methods exhibited strong correlation (r: 0.78) but poor agreement (CCC: 0.34), though agreement increased when time in PA was averaged across all valid days (r: 0.92, CCC: 0.71) (Table 4). Trends were consistent across disease severity strata (S3 Table).

2. Time in PA: Did Fitbit agree with criterion measures?. During scripted tasks, epoch-level agreement between Fitbit, manual counts, and the two Actigraph methods was excellent (k: 0.85–0.87) (Table 4). During free-living evaluation, agreement was moderate to excellent at the epoch level (k: 0.62–0.76), poor at the daily level (CCC: 0.18–0.36) and weak to moderate at the average level (CCC:0.35–0.52). Fitbit consistently exhibited higher agreement with the

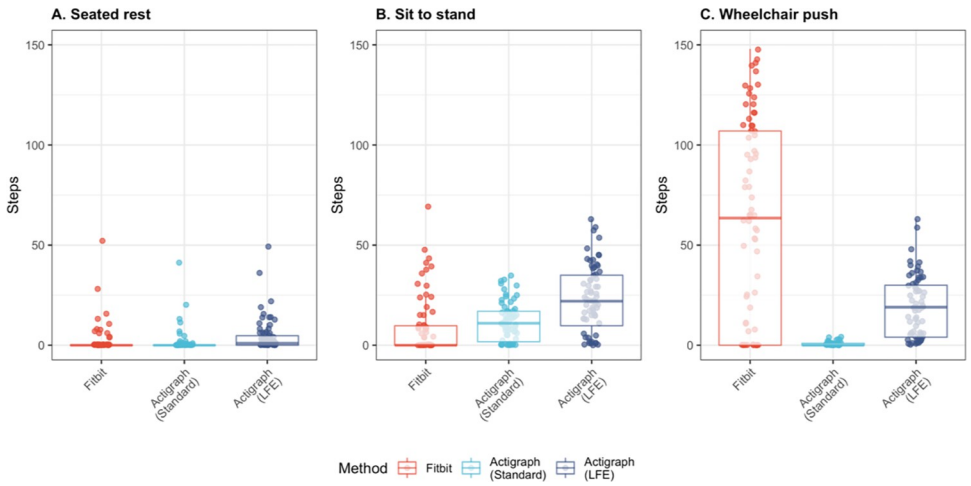


Fig 4. Steps per minute derived from Fitbit and Actigraph during scripted non-walking tasks. Step counts were generated by each method–Fitbit, Actigraph (Standard), and Actigraph (LFE)–during tasks in which no manual steps were identified: A) Seated at rest, B) during the sit to stand task, and C) while pushing a wheelchair.

https://doi.org/10.1371/journal.pdig.0000171.g004

Actigraph (Vertical) method than the Actigraph (VM) method (Table 4). However, both Actigraph methods registered significantly more time in PA than Fitbit during free-living PA and limits of agreement on Bland Altman plots were wide for all pairs of measures (Fig 6). Agreement was consistent across subgroups with mild and moderate MS, but was consistently lower in the subgroup with severe MS (S3 Table).

3. Time in PA: Did Fitbit correlate with criterion measures?. During free-living evaluation, correlations between Fitbit and Actigraph-derived time in PA were moderate to strong at the daily (r: 0.74–0.82) and average (r: 0.72–0.76) levels (Table 4). Correlations at the daily level were consistent across subgroups (r: 0.70–0.84), but those at the average level were lower in the subgroup with severe MS (Mild–r: 0.87–0.89; Moderate–r: 0.83–0.87; Severe–r:0.38–0.48)(S3 Table).

4. Time in PA: Did Fitbit associate with clinical outcomes?. Fitbit consistently exhibited moderate to strong correlations with clinical outcome measures, though no method exhibited differences between mild and moderate or severe MS (Table 3). These relationships were either similar to or stronger than those exhibited by Actigraph-derived PA metrics.

Time in moderate to vigorous physical activity

1. Time in MVPA: Did criterion measures agree and correlate with each other?. The three Actigraph methods–Actigraph (Uniform), Actigraph (Severity), and Actigraph (Sasaki)–exhibited excellent pairwise agreement at the epoch level during scripted tasks (k: 0.76–0.84) and good to excellent agreement during free-living evaluation (k: 0.68–0.82) (Table 4). At the daily level, the three methods exhibited strong pairwise correlations (r: 0.78–0.82) and good pairwise agreement (CCC: 0.56–0.63) (Table 4). Correlations and agreement further increased when time in PA was averaged across all valid days (r: 0.76–0.90, CCC: 0.72–0.88) (Table 4).

Trends were not consistent across disease severity strata. Correlation and agreement between Actigraph methods were consistently lower in persons with severe MS compared to those with mild or moderate MS. Correlation and agreement also tended to be higher in the subgroup with moderate MS compared to that with mild MS, though this was not consistent across all levels of data aggregation. (S4 Table).

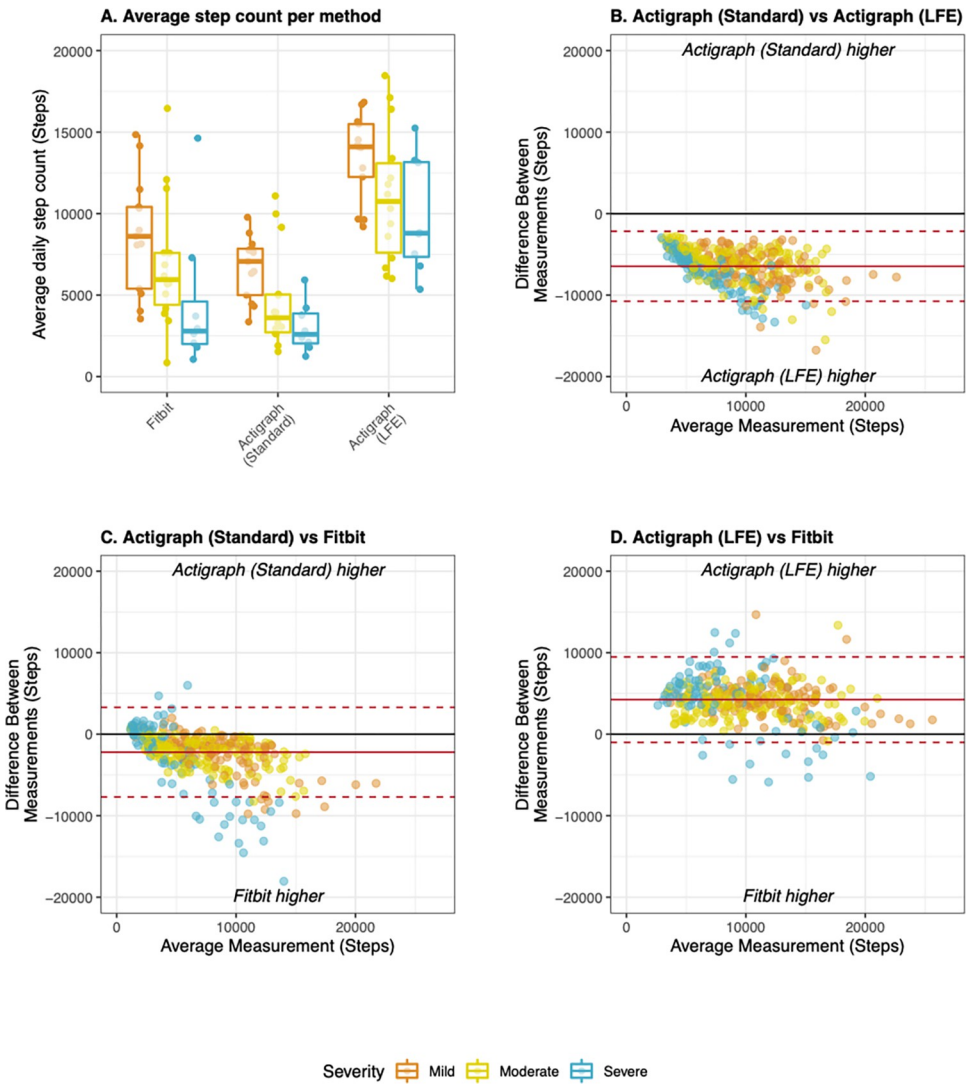


Fig 5. Step counts derived from Fitbit and Actigraph during free-living evaluation. A) Average step counts derived from the Fitbit, Actigraph (Standard), and Actigraph (LFE) stratified by disease severity. B) Bland Altman plot comparing Actigraph methods to each other; C) Bland Altman plot comparing Fitbit to Actigraph (Standard), D) Bland Altman plot comparing Fitbit to Actigraph (LFE). All Bland Altman plots display data collected during rehab and at home at the daily level of aggregation. Mean bias (solid line) and limits of agreement (dotted lines) were adjusted for patient-level random effects. Mild MS: EDSS < 4.0, Moderate MS: EDSS 4.0–5.5, Severe MS: EDSS 6.0–6.5.

https://doi.org/10.1371/journal.pdig.0000171.g005

2. MVPA: Did Fitbit agree with criterion measures?. During scripted tasks, Fitbit and Actigraph exhibited no agreement in their categorization of MVPA (Table 4). The Fitbit rarely classified any scripted task as MVPA, whereas Actigraph methods frequently classified the 6MWT, Sit to Stand, Stair Climbing, and Walking with Postural Transitions tasks as MVPA. It remains unclear which activities register as MVPA on the Fitbit in this population.

During free-living evaluations, overall pairwise agreement between Fitbit and Actigraph methods was poor to fair at the epoch (k: 0.39–0.42) level, fair at the daily level (CCC: 0.41–0.45), and good to excellent at the average level (CCC: 0.63–0.80) (Table 4). Bland Altman analysis showed that median bias between methods was low in all cases, though limits of agreement between methods were wider when Fitbit was compared to Actigraph methods than

Table 3. Correlations between device-derived physical activity metrics and clinical measures.

	MSWS-12	IPAQ	EDSS	6MWT	10mGS	Mild vs Moderate/Severe (Based on EDSS)	
	r [95% CI]	r [95% CI]	r [95% CI]	r [95% CI]	r [95% CI]	p	ES [95% CI]
Step count							
Fitbit	-0.52 [-0.73 - -0.22]	0.92 [0.66–0.98]	-0.50 [-0.72 - -0.20]	0.62 [0.31–0.81]	-0.47 [-0.70 - -0.16]	0.0358	0.36 [0.05–0.62]
Act (Stand)	-0.67 [-0.82 - -0.43]	0.90 [0.57–0.98]	-0.64 [-0.80 - -0.39]	0.72 [0.47–0.86]	-0.65 [-0.81 - -0.40]	0.0022	0.52 [0.23–0.75]
Act (LFE)	-0.56 [-0.76 - -0.28]	0.76 [0.20–0.95]	-0.48 [-0.70 - -0.17]	0.58 [0.25–0.78]	-0.61 [-0.78 - -0.34]	0.0194	0.40 [0.11–0.65]
Physical activity							
Fitbit	-0.51 [-0.72 - -0.20]	0.93 [0.71–0.99]	-0.48 [-0.70 - -0.18]	0.52 [0.17–0.75]	-0.40 [-0.65 - -0.07]	0.0876	0.30 [0.02–0.57]
Act (Vert)	-0.45 [-0.69 - -0.14]	0.77 [0.21–0.95]	*-0.27 [-0.55–0.07]*	*0.36 [-0.02–0.65]*	-0.43 [-0.67 - -0.11]	0.0942	0.29 [0.02–0.59]
Act (VM)	-0.42 [-0.66 - -0.09]	0.69 [0.05–0.93]	*-0.26 [-0.55–0.08]*	*0.28 [-0.11–0.60]*	-0.40 [-0.65 - -0.08]	0.1521	0.25 [0.01–0.55]
Moderate to vigorous physical activity							
Fitbit	-0.42 [-0.66 - -0.09]	0.83 [0.36–0.96]	-0.37 [-0.63 - -0.05]	0.54 [0.20–0.76]	-0.45 [-0.68 - -0.13]	0.0508	0.34 [0.04–0.63]
Act (Uni)	-0.56 [-0.76 - -0.28]	*0.62 [-0.07–0.91]*	-0.51 [-0.72 - -0.21]	0.63 [0.33–0.82]	-0.51 [-0.72 - -0.20]	0.0004	0.60 [0.34–0.79]
Act (Sev)	-0.46 [-0.69 - -0.14]	*0.49 [-0.26–0.87]*	*-0.23 [-0.52–0.11]*	0.51 [0.15–0.74]	*-0.31 [-0.59–0.03]*	0.0097	0.45 [0.15–0.68]
Act (Sasaki)	*-0.32 [-0.59–0.02]*	*0.26 [-0.49–0.79]*	-0.35 [-0.62 - -0.02]	*0.38 [0.00–0.66]*	*-0.33 [-0.60–0.01]*	0.0284	0.38 [0.08–0.63]

Point estimates which did not reach statistical significance, defined here as the 95% confidence intervals excluding 0, are shown in italics.

Act: Actigraph; Stand: Standard; LFE: Low frequency extension; Vert: Vertical; VM: Vector Magnitude; Uni: Uniform; Sev: Severity; ES: effect size; r: Pearson's correlation coefficient

https://doi.org/10.1371/journal.pdig.0000171.t003

when Actigraph methods were compared to each other (Fig 7). Agreement differed across disease severity strata at all levels of aggregation. Agreement was consistently highest in those with moderate MS, slightly lower in those with mild MS, and lowest in those with severe MS (S4 Table).

3. Time in MVPA: Did Fitbit correlate with criterion measures?. During free-living evaluation, correlations between Fitbit and Actigraph methods were moderate at the daily

Table 4. Correlation and agreement between time in physical activity and moderate to vigorous physical activity derived from Fitbit and Actigraph.

	Epoch level, Scripted tasks	Epoch level, Free living	Daily level, Free living		Average level, Free living	
Comparison	k [95% CI]	k [95% CI]	r [95% CI]	CCC [95% CI]	r [95% CI]	CCC [95% CI]
Physical Activity						
Act(VM) vs Act(Vert) [a]	0.93 [0.88–0.97]	0.75 [0.73–0.77]	0.78 [0.74–0.83]	0.34 [0.25–0.41]	0.92 [0.85–0.96]	0.71 [0.57–0.81]
Fitbit vs Act(Vert)	0.87 [0.72–1.02]	0.76 [0.73–0.80]	0.74 [0.67–0.80]	0.36 [0.21–0.50]	0.72 [0.50–0.85]	0.52 [0.32–0.68]
Fitbit vs Act(VM)	0.85 [0.71–0.99]	0.62 [0.58–0.66]	0.82 [0.77–0.86]	0.18 [0.11–0.26]	0.76 [0.57–0.87]	0.35 [0.20–0.49]
Moderate to Vigorous Physical Activity						
Act(Uni) vs Act(Sev) [a]	0.84 [0.77–0.91]	0.82 [0.75–0.89]	0.79 [0.68–0.86]	0.63 [0.46–0.74]	0.89 [0.79–0.94]	0.87 [0.77–0.93]
Act(Uni) vs Act(Sasaki) [a]	0.79 [0.63–0.94]	0.73 [0.66–0.80]	0.82 [0.72–0.88]	0.63 [0.47–0.73]	0.90 [0.82–0.95]	0.88 [0.78–0.94]
Act(Sev) vs Act(Sasaki) [a]	0.76 [0.63–0.88]	0.68 [0.60–0.76]	0.78 [0.65–0.85]	0.56 [0.38–0.66]	0.76 [0.57–0.87]	0.72 [0.53–0.84]
Fitbit vs Act(Uni)	*-0.18 [-0.48–0.13]*	0.39 [0.27–0.51]	0.68 [0.56–0.80]	0.45 [0.27–0.63]	0.80 [0.64–0.90]	0.80 [0.64–0.89]
Fitbit vs Act(Sev)	*-0.18 [-0.48–0.13]*	0.42 [0.31–0.54]	0.65 [0.55–0.75]	0.41 [0.27–0.56]	0.64 [0.40–0.81]	0.63 [0.39–0.79]
Fitbit vs Act(Sasaki)	*-0.18 [-0.46–0.10]*	0.41 [0.31–0.52]	0.67 [0.54–0.79]	0.44 [0.27–0.58]	0.79 [0.62–0.89]	0.78 [0.61–0.88]

[a] Comparison between two criterion measures

Point estimates which did not reach statistical significance, defined here as the 95% confidence intervals excluding 0, are shown in italics.

Act: Actigraph; Vert: Vertical; VM: Vector Magnitude; Uni: Uniform; Sev: Severity; k: Fleiss' kappa; CI: confidence interval; r: Pearson correlation coefficient; CCC: Lin's Concordance correlation coefficient

https://doi.org/10.1371/journal.pdig.0000171.t004

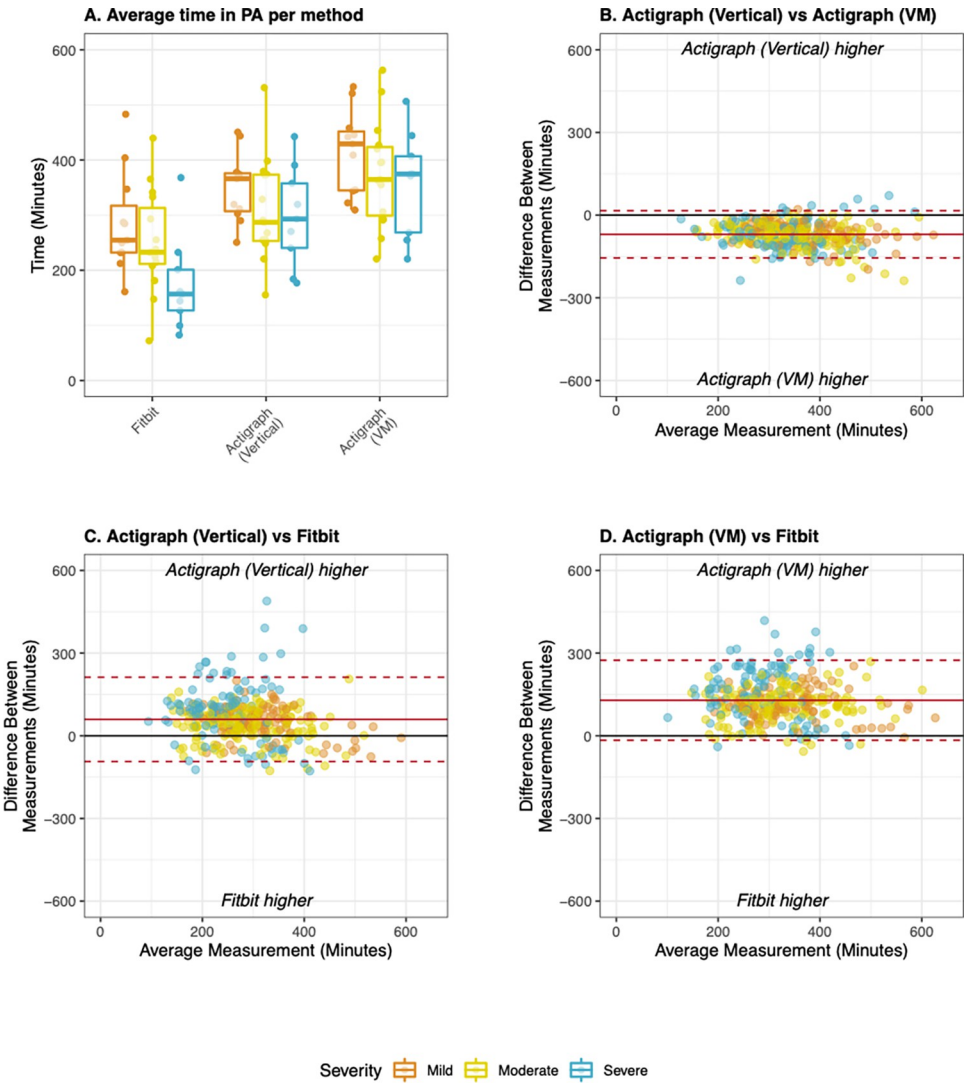


Fig 6. Total time in physical activity derived from Fitbit and Actigraph during free-living evaluation. A) Average time in PA derived from the Fitbit, Actigraph (Vertical), and Actigraph (VM) stratified by disease severity; B) Bland Altman plot comparing both Actigraph methods to each other; C) Bland Altman plot comparing Fitbit to Actigraph (Vertical), D) Bland Altman plot comparing Fitbit to Actigraph (VM). All Bland Altman plots display data collected during rehab and at home at the daily level of aggregation. Mean bias (solid line) and limits of agreement (dotted lines) were adjusted for patient-level random effects. Mild MS: EDSS < 4.0, Moderate MS: EDSS 4.0–5.5, Severe MS: EDSS 6.0–6.5.

https://doi.org/10.1371/journal.pdig.0000171.g006

level (r: 0.65–0.68) and moderate to strong at the average level (r: 0.64–0.80) (Table 4). Correlations were highest in the subgroup with moderate MS, lower in the subgroup with mild MS, and lowest in the subgroup with severe MS (S4 Table).

4. Time in MVPA: Did Fitbit associate with clinical outcomes?. Fitbit-derived time in MVPA exhibited moderate to strong correlations with clinical outcome measures, whereas the Actigraph methods often did not (Table 3). Fitbit exhibited the expected differences between groups based on MSWS-12 cutoffs, but differences between mild/ moderate and severe MS did not reach statistical significance. All Actigraph-derived PA estimates exhibited differences between subgroups (Table 3).

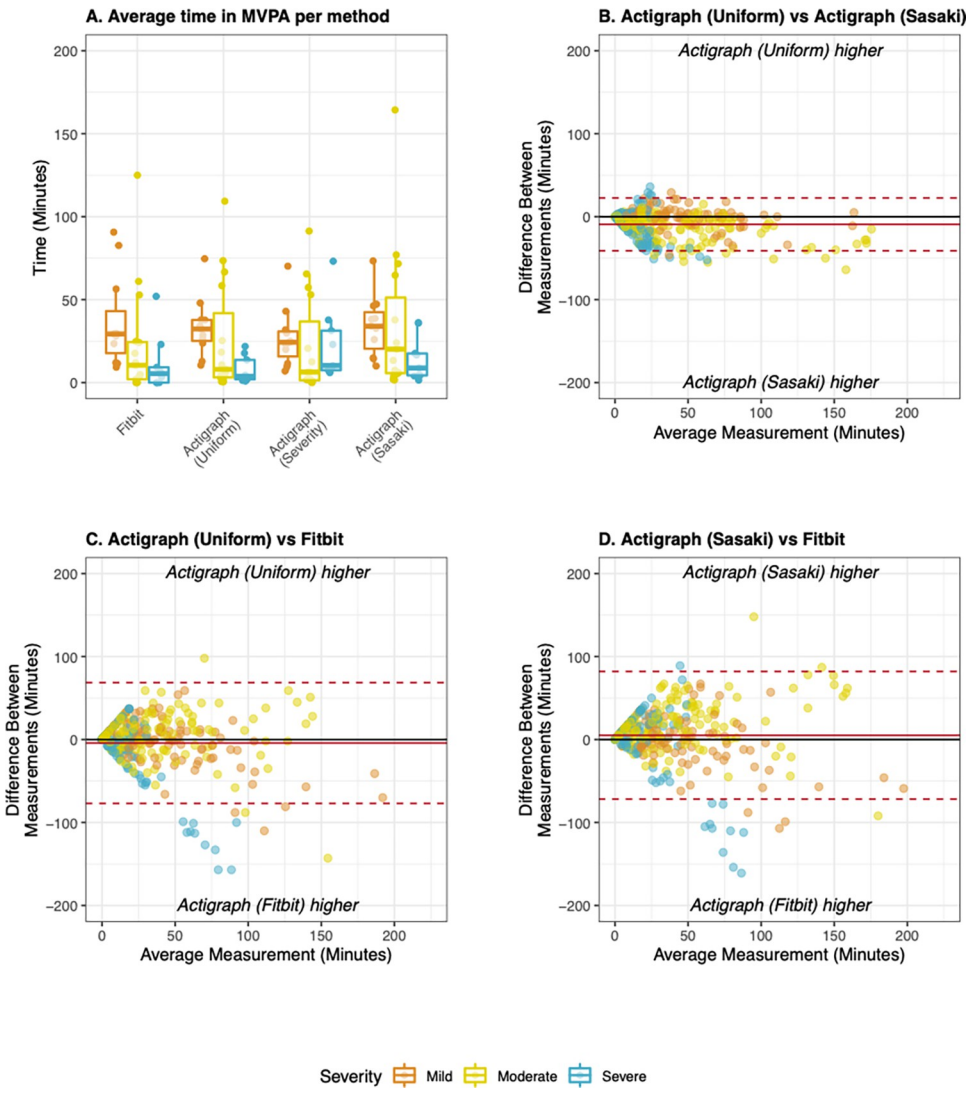


Fig 7. Time in moderate to vigorous physical activity derived from Fitbit and Actigraph during free living evaluation. A) Average time in PA derived from the Fitbit, Actigraph (Uniform), and Actigraph (Sasaki) stratified by disease severity. B) Bland Altman plot comparing both Actigraph methods to each other; C) Bland Altman plot comparing Fitbit to Actigraph (Uniform), D) Bland Altman plot comparing Fitbit to Actigraph (Sasaki). All Bland Altman plots display data collected during rehab and at home at the daily level of aggregation. Mean bias (solid line) and limits of agreement (dotted lines) were adjusted for patient-level random effects. Mild MS: EDSS < 4.0, Moderate MS: EDSS 4.0–5.5, Severe MS: EDSS 6.0–6.5.

https://doi.org/10.1371/journal.pdig.0000171.g007

Triangulating the validity of Fitbit-derived PA metrics

Qualitative ratings generated through the triangulation process are shown in Table 5. Fitbit and Actigraph-derived PA metrics cannot be considered equivalent. Nor can most Actigraph methods be considered equivalent to each other. However, all measures exhibited evidence of construct validity.

Discussion

In this study, we evaluated the validity of PA metrics derived from the Fitbit Inspire HR during scripted walking tasks and free-living activity at multiple levels of data aggregation. Fitbit-

Table 5. Triangulating the validity of Fitbit-derived PA metrics.

	Step count	Time in PA	Time in MVPA
***Did criterion measures agree with each other*?**			
Scripted tasks	+	++	++
Free living, Epoch level	–	++	+/++
Free living, Daily level	–	+	+
Free living, Average level	-	-	++
Consistency across severity strata*	+	+	-
***Did criterion measures correlate with each other*?**			
Scripted tasks	++	na	na
Free living, Epoch level	-	na	na
Free living, Daily level	++	++	++
Free living, Average level	++	++	++
Consistency across severity strata*	+	+	-
***Did Fitbit agree with criterion measures*?**			
Scripted tasks	+	++	–
Free living, Epoch level	–	+/++	-/+
Free living, Daily level	-/+	–/-	+
Free living, Average level	+	-/+	+/++
Consistency across severity strata*	-	-	-
Did Fitbit meet or exceed the agreement exhibited by criterion measures?	+	-/+	-
***Did Fitbit correlate with criterion measures*?**			
Scripted tasks	++	na	na
Free living, Epoch level	-	na	na
Free living, Daily level	++	+/++	+
Free living, Average level	++	+/++	+/++
Consistency across severity strata*	-/+	-	-
Did Fitbit meet or exceed the correlations exhibited by criterion measures?	+	-/+	-
***Did Fitbit relate to clinical outcomes*?**			
Did Fitbit exhibit the expected correlations with clinical measures?	+	+	+
Did Fitbit-derived PA metrics differ across known groups?	+	-	+
Did Fitbit meet or exceed the relationships exhibited by criterion measures?	+	+	+
Validity of Fitbit-derived PA metrics			
Can criterion measures be considered equivalent?	-	-	-
Can Fitbit and Actigraph be considered equivalent?	-	-	-
Did Fitbit exhibit evidence of construct validity?	+	+	+

++: Excellent agreement or strong correlations (0.75–1.0)

+: Fair to good agreement or moderate correlations (0.4–0.75)

-: Poor agreement, weak correlations (0.2–0.4)

—: Very weak or complete lack of agreement or correlation (<0.2)

+/-: Evidence was mixed; Binary yes/no responses are indicated with + or -, respectively

*Consistency across severity strata describes whether the trends observed during scripted tasks and at the epoch, daily, and average levels were consistent across subgroups with mild, moderate, and severe MS. A positive rating in this category does not necessarily mean that measures correlated or agreed with each other.

https://doi.org/10.1371/journal.pdig.0000171.t005

derived PA metrics demonstrated construct validity, but not equivalency with criterion measures derived from the Actigraph GT3X. Correlations and agreement between measures differed across settings, data aggregation levels, and disease severity strata. However, criterion measures exhibited limited agreement amongst themselves, and we demonstrate that, in most cases, Fitbit performs within the range of their inter-method variability. In light of these findings, consumer-grade fitness trackers such as Fitbit may be advantageous for long-term PA tracking in PwMS.

Evaluating the validity of Fitbit-derived physical activity metrics

Step count. Our triangulation suggests that Fitbit-derived step count may outperform Actigraph-derived step count during free-living PA in people with mild or moderate MS, but should be used with caution in those with severe walking impairment. We found that Fitbit-derived step counts exhibited strong correlations but poor agreement with Actigraph-derived step counts. This is consistent with previous studies comparing Fitbit-derived step counts to those derived from Actigraph in healthy populations [77–80] and MS populations [30,31]. However, Actigraph-derived step counts often exhibited worse agreement with each other than they did with the Fitbit, and step counts derived from the Actigraph (LFE) were considered unrealistically high for this population by clinical experts (authors JK, RG). This pattern is consistent with a previous investigation, in which Actigraph (Standard) underestimated step count by 25–30%, and Actigraph (LFE) overestimated step count by 30% [33]. Therefore the 'true' step count likely falls somewhere between these two metrics, as Fitbit-derived step counts did in this study. Fitbit demonstrated different sources of error than Actigraph methods during scripted tasks which disproportionately impacted persons with severe MS. This reduced performance has been previously attributed to reductions in walking speed [59]. However, our observations suggest that wheelchair or rollator use and common balance management strategies such as holding on to furniture for support [81,82] affect upper body movement while walking and may contribute to reduced performance in those with severe MS.

Time in physical activity. Fitbit-derived time in PA was consistently lower than both Actigraph methods, though the three correlated strongly and exhibited moderate to strong agreement at most levels of data aggregation. Neither Actigraph method has been validated under free-living conditions. Therefore, it is unclear how they relate to "true" time in free-living physical activity, defined as "any voluntary bodily movement produced by the skeletal muscles that requires energy expenditure" by the World Health Organization [83]. We offer three potential explanations for the differences between Actigraph and Fitbit-derived time in PA. First, Fitbit's sensitivity to PA may simply be reduced in populations with MS, as PwMS have altered gait compared to healthy controls [26,27]. Alternatively, Fitbit's mischaracterization of activity related to upper rather than lower body motion may introduce different biases into Fitbit and Actigraph-derived measurements, as it did for step count. Finally, the inclusion of heart rate in Fitbit's PA algorithms may yield measurements of a highly related, but slightly different PA construct than that measured by Actigraph. Nevertheless, Fitbit exhibited greater evidence of construct validity in this study (Table 3). The criterion validity of all three methods should be confirmed in future work.

Time in moderate to vigorous physical activity. Fitbit-derived MVPA exhibited no agreement with Actigraph-derived MVPA during scripted tasks and only poor to fair agreement at the epoch or daily level. Previous studies in healthy populations similarly suggest disagreement between the Fitbit the Actigraph. Two recent systematic reviews of Fitbit validation studies suggest that, in healthy populations, Fitbit strongly correlates with Actigraph-derived MVPA [19], though point estimates of MVPA derived from the Fitbit overestimate time in

MVPA compared to the Actigraph [20]. On the surface, our findings align with these findings. However, in this study, Fitbit-derived MVPA exhibited evidence of convergent and known-groups validity, whereas Actigraph-derived PA metrics often did not. This, and the fact that Fitbit did not register scripted activities as MVPA, suggest that Fitbit-derived MVPA reflects a different construct than Actigraph-derived MVPA. The differences between these constructs could not be characterized based on the evidence generated in this study, but may relate to the inclusion of heart rate and upper body PA in Fitbit's activity intensity assessment algorithms.

The case for consumer-grade activity monitors

If Fitbit-derived metrics do indeed exhibit an acceptable level of construct validity in PwMS, they present new opportunities as long-term, engaging, and user-friendly PA monitoring tools. Current monitoring practices rely on questionnaires, diaries, or research-grade wearable devices. However, the user experiences and validity of these methods are also limited. Questionnaires and diaries are burdensome to complete regularly, subject to recall bias, and insensitive to short bouts of light or lifestyle physical activity [84–87]. Participants in previous studies have reported that research-grade wearable devices are "bulky," "uncomfortable," and "attract unwanted attention" during free-living PA tracking [88]. Further, PwMS prefer to receive feedback about their activity from devices that they wear during studies [89,90], which the Actigraph and many other research-grade accelerometers do not provide [88]. Conversely, Fitbit devices are considered comfortable and inconspicuous [91]. They collect data passively and provide regular feedback to the wearer, potentially increasing long-term engagement with PA monitoring [17,92]. We provisionally demonstrated this effect in this study, as Fitbit wear time was higher than that of the Actigraph. Finally, the Fitbit Inspire HR is relatively affordable compared to research grade devices at a price of approximately 100 US dollars [93]. Fitbit-derived PA metrics may not be fit for all research purposes, for example as outcome measures in efficacy studies which may be confounded by the device's feedback. Nevertheless, the rich, longitudinal data derived from Fitbit devices could reveal novel insights and patterns not discoverable through current PA assessments.

Strengths, limitations, and future work

This study investigated the validity of Fitbit-derived physical activity metrics according to best practices, accounting for known shortcomings of widely-used reference measurements. It explored the construct validity of three Fitbit-derived PA metrics in a systematic manner, in multiple settings, patient subgroups, and levels of data aggregation. It therefore represents, to the authors' knowledge, the most comprehensive evaluation of a Fitbit's validity to date.

However, this study is not without its limitations. Johnston et al. recommend video monitoring as a criterion measure for step counts during free-living evaluation, though they note that this method is frequently infeasible due to processing time and patient burden [32]. We opted not to use video during free-living evaluation, instead addressing the known shortcomings of available criterion methods through triangulation [57]. Similarly, we did not use calorimetry to derive PA intensity during any tasks, as these are difficult and burdensome to implement as criterion measures in contexts other than scripted walking. We therefore cannot quantify the Fitbit's absolute accuracy through the present study, and our findings should be considered relative to the known benefits and shortcomings of Actigraph methods.

The findings presented here are necessary, but not sufficient, to support the use of Fitbit-derived PA metrics for MS. If Fitbit-derived metrics are to be used to self-manage PA, track PA over time, or evaluate the efficacy of novel interventions, they must be able to detect change at the patient and population level. It is possible that the biases demonstrated here could

impact Fitbit-derived metrics' ability to detect change on both the individual and population level, and future work should evaluate their responsiveness. Novel analysis methods which can account for these confounding effects, especially those which capitalize on the richness of long-term PA data, should also be the subject of future research.

Conclusions

Fitbit-derived metrics are not equivalent to those derived from Actigraph. However, they exhibit similar or stronger evidence of construct validity. Consumer-grade fitness trackers such as the Fitbit may therefore be suitable as PA management tools for people with mild or moderate MS–particularly to monitor intra-individual temporal changes. However, they should be used with caution in populations with advanced walking impairment. Future work should investigate the criterion validity and responsiveness of both Fitbit and Actigraph-derived PA metrics.

Supporting information

S1 Text. Standard operating procedure: Manual step counts.
(DOCX)

S1 Fig. Epoch-level agreement between Fitbit-derived and Actigraph-derived step count. A) between Fitbit and Actigraph (Standard), B) between Fitbit and Actigraph (LFE), and C) between both Actigraph methods. Each point represents the number of steps counted in a 5-minute epoch. Epochs in perfect agreement fall along the black diagonal line. The majority of epochs were relatively consistent, though not in perfect agreement, and fell near the diagonal. A relatively small portion of epochs exhibited very low counts by one device and high counts by the other device. Two percent (Standard filter, panel A) and four percent (LFE, panel B) of exhibited this distribution. This discrepancy may be related to the limited specificity of the wear time algorithm, the limited ability of the Actigraph (Standard) method to detect impaired gait, or the differing sources of bias between devices the devices. In panel C, all points are above the diagonal because the LFE always yielded step counts which were equal to or greater than those with the standard filter.
(DOCX)

S1 Table. Participant characteristics during each stage of the validity evaluation.
(DOCX)

S2 Table. Correlation and agreement between step counts derived from Fitbit and Actigraph.
(DOCX)

S3 Table. Correlation and agreement between time in physical activity derived from Fitbit and Actigraph.
(DOCX)

S4 Table. Correlation and agreement between time in moderate to vigorous physical activity derived from Fitbit and Actigraph.
(DOCX)

Author Contributions

Conceptualization: Ashley Polhemus, Chloé Sieber, Christina Haag, Jan Kool, Roman Gonzenbach, Viktor von Wyl.

Data curation: Ashley Polhemus, Chloé Sieber, Christina Haag, Ramona Sylvester.

Formal analysis: Ashley Polhemus.

Investigation: Chloé Sieber, Christina Haag, Ramona Sylvester, Viktor von Wyl.

Methodology: Ashley Polhemus, Chloé Sieber, Christina Haag, Jan Kool, Viktor von Wyl.

Project administration: Ramona Sylvester.

Resources: Ramona Sylvester.

Software: Ashley Polhemus.

Supervision: Jan Kool, Roman Gonzenbach, Viktor von Wyl.

Validation: Ashley Polhemus.

Visualization: Ashley Polhemus.

Writing – original draft: Ashley Polhemus.

Writing – review & editing: Ashley Polhemus, Chloé Sieber, Christina Haag, Ramona Sylvester, Jan Kool, Roman Gonzenbach, Viktor von Wyl.

References

1. Beckerman H, de Groot V, Scholten MA, Kempen JCE, Lankhorst GJ. Physical activity behavior of people with multiple sclerosis: understanding how they can become more physically active. Phys Ther. 2010; 90: 1001–1013. https://doi.org/10.2522/ptj.20090345 PMID: 20508028
2. Kalb R, Brown TR, Coote S, Costello K, Dalgas U, Garmon E, et al. Exercise and lifestyle physical activity recommendations for people with multiple sclerosis throughout the disease course. Mult Scler J. 2020; 26: 1459–1469. https://doi.org/10.1177/1352458520915629 PMID: 32323606
3. Razazian N, Kazeminia M, Moayedi H, Daneshkhah A, Shohaimi S, Mohammadi M, et al. The impact of physical exercise on the fatigue symptoms in patients with multiple sclerosis: A systematic review and meta-analysis. BMC Neurol. 2020; 20: 93. https://doi.org/10.1186/s12883-020-01654-y PMID: 32169035
4. Motl RW, Pilutti LA. The benefits of exercise training in multiple sclerosis. Nature Reviews Neurology. Nature Publishing Group; 2012. pp. 487–497. https://doi.org/10.1038/nrneurol.2012.136 PMID: 22825702
5. Motl RW, Sandroff BM, Kwakkel G, Dalgas U, Feinstein A, Heesen C, et al. Exercise in patients with multiple sclerosis. Lancet Neurol. 2017; 16: 848–856. https://doi.org/10.1016/S1474-4422(17)30281-8 PMID: 28920890
6. Kasser S. Exercising with multiple sclerosis: Insights into meaning and motivation. Adapt Phys Act Q. 2009; 26: 274–289. https://doi.org/10.1123/apaq.26.3.274 PMID: 19799098
7. Learmonth YC, Motl RW. Physical activity and exercise training in multiple sclerosis: A review and content analysis of qualitative research identifying perceived determinants and consequences. Disabil Rehabil. 2016; 38: 1227–1242. https://doi.org/10.3109/09638288.2015.1077397 PMID: 26314587
8. Aminian S, Ezeugwu VE, Motl RW, Manns PJ. Sit less and move more: perspectives of adults with multiple sclerosis. Disabil Rehabil. 2019; 41: 904–911. https://doi.org/10.1080/09638288.2017.1416499 PMID: 29262734
9. Adamson BC, Adamson MD, Littlefield MM, Motl RW. 'Move it or lose it': perceptions of the impact of physical activity on multiple sclerosis symptoms, relapse and disability identity. Qual Res Sport Exerc Heal. 2017; 10: 457–475. https://doi.org/10.1080/2159676X.2017.1415221
10. Kayes NM, McPherson KM, Taylor D, Schluter PJ, Kolt GS. Facilitators and barriers to engagement in physical activity for people with multiple sclerosis: a qualitative investigation. Disabil Rehabil. 2011; 33: 625–642. https://doi.org/10.3109/09638288.2010.505992 PMID: 20695816
11. Giunti G, Kool J, Rivera Romero O, Dorronzoro Zubiete E. Exploring the Specific Needs of Persons with Multiple Sclerosis for mHealth Solutions for Physical Activity: Mixed-Methods Study. JMIR mHealth uHealth. 2018; 6: e37. https://doi.org/10.2196/mhealth.8996 PMID: 29426814

12. Block VJ, Alexander AM, Papinutto N, Rajesh A, Gundel T, Gelfand JM, et al. Remotely monitored ambulatory activity correlates with disability in progressive MS: Baseline data from the SPI2 PH3 trial of MD1003 (HDPB BIOTIN). Mult Scler J. 2020; 26: 34–35. https://doi.org/10.1177/1352458520917096

13. Zabalza A, Guerrero AI, Buron M, Dalla Costa G, La Porta ML, Martinis M, et al. Descriptive study on recruitment effort for a remote monitoring study in multiple sclerosis: Radar study. Mult Scler J. 2020; 26: 146–147. https://doi.org/10.1177/1352458520974937

14. Cree BAC, Cutter G, Wolinsky JS, Freedman MS, Comi G, Giovannoni G, et al. Safety and efficacy of MD1003 (high-dose biotin) in patients with progressive multiple sclerosis (SPI2): a randomised, double-blind, placebo-controlled, phase 3 trial. Lancet Neurol. 2020; 19: 988–997. https://doi.org/10.1016/S1474-4422(20)30347-1 PMID: 33222767

15. Latimer-Cheung AE, Martin Ginis KA, Hicks AL, Motl RW, Pilutti LA, Duggan M, et al. Development of evidence-informed physical activity guidelines for adults with multiple sclerosis. Archives of Physical Medicine and Rehabilitation. W.B. Saunders; 2013. pp. 1829–1836.e7. https://doi.org/10.1016/j.apmr.2013.05.015 PMID: 23770262

16. Simblett SK, Evans J, Greer B, Curtis H, Matcham F, Radaelli M, et al. Engaging across dimensions of diversity: A cross-national perspective on mHealth tools for managing relapsing remitting and progressive multiple sclerosis. Mult Scler Relat Disord. 2019; 32: 123–132. https://doi.org/10.1016/j.msard.2019.04.020 PMID: 31125754

17. Simblett S, Greer B, Matcham F, Curtis H, Polhemus A, Ferrao J, et al. Barriers and facilitators to engagement with remote measurement technology for managing health: a systematic review and content analysis of findings. J Med Internet Res. 2018; 20: e10480. https://doi.org/10.2196/10480 PMID: 30001997

18. Irwin C, Gary R. Systematic Review of Fitbit Charge 2 Validation Studies for Exercise Tracking. Transl J Am Coll Sport Med. 2022; 7: 1–7. https://doi.org/10.1249/TJX.0000000000000215

19. Gorzelitz J, Farber C, Gangnon R, Cadmus-Bertram L. Accuracy of Wearable Trackers for Measuring Moderate- to Vigorous-Intensity Physical Activity: A Systematic Review and Meta-Analysis. J Meas Phys Behav. 2020; 3: 346–357. https://doi.org/10.1123/jmpb.2019-0072

20. Feehan LM, Geldman J, Sayre EC, Park C, Ezzat AM, Young Yoo J, et al. Accuracy of fitbit devices: Systematic review and narrative syntheses of quantitative data. JMIR mHealth and uHealth. JMIR Publications Inc.; 2018. p. e10527. https://doi.org/10.2196/10527 PMID: 30093371

21. Straiton N, Alharbi M, Bauman A, Neubeck L, Gullick J, Bhindi R, et al. The validity and reliability of consumer-grade activity trackers in older, community-dwelling adults: A systematic review. Maturitas. Elsevier Ireland Ltd; 2018. pp. 85–93. https://doi.org/10.1016/j.maturitas.2018.03.016 PMID: 29704922

22. Wong CK, Mentis HM, Kuber R. The bit doesn't fit: Evaluation of a commercial activity-tracker at slower walking speeds. Gait Posture. 2018; 59: 177–181. https://doi.org/10.1016/j.gaitpost.2017.10.010 PMID: 29049964

23. Schaffer SD, Holzapfel SD, Fulk G, Bosch PR. Step count accuracy and reliability of two activity tracking devices in people after stroke. Physiother Theory Pract. 2017; 33: 788–796. https://doi.org/10.1080/09593985.2017.1354412 PMID: 28777710

24. Katzan I, Schuster A, Kinzy T. Physical activity monitoring using a fitbit device in ischemic stroke patients: Prospective cohort feasibility study. JMIR mHealth and uHealth. JMIR Publications Inc.; 2021. p. e14494. https://doi.org/10.2196/14494 PMID: 33464213

25. Floegel TA, Florez-Pregonero A, Hekler EB, Buman MP. Validation of Consumer-Based Hip and Wrist Activity Monitors in Older Adults With Varied Ambulatory Abilities. Journals Gerontol Ser A Biol Sci Med Sci. 2017; 72: 229–236. https://doi.org/10.1093/gerona/glw098 PMID: 27257217

26. Polhemus A, Ortiz LD, Brittain G, Chynkiamis N, Salis F, Gaßner H, et al. Walking on common ground: a cross-disciplinary scoping review on the clinical utility of digital mobility outcomes. npj Digit Med. 2021;4. https://doi.org/10.1038/s41746-021-00513-5 LK - https://uzb.swisscovery.slsp.ch/openurl/41SLSP_UZB/41SLSP_UZB:UZB?sid=Elsevier:EMBASE&sid=EMBASE&issn=23986352&id=doi:10.1038%2Fs41746-021-00513-5&atitle=Walking+on+common+ground%3A+a+cross-disciplinary+scoping+review+on+the+clinical+utility+of+digital+mobility+outcomes&stitle=npj+Digit.+Med.&title=npj+Digital+Medicine&volume=4&issue=1&spage=&epage=&aulast=Polhemus&aufirst=Ashley&auinit=A.&aufull=Polhemus+A.&coden=&isbn=&pages=-&date=2021&auinit1=A&auinitm=

27. Comber L, Galvin R, Coote S. Gait deficits in people with multiple sclerosis: A systematic review and meta-analysis. Gait Posture. 2017; 51: 25–35. https://doi.org/10.1016/j.gaitpost.2016.09.026 PMID: 27693958

28. Squires LA, Williams N, Morrison VL. Matching and accepting assistive technology in multiple sclerosis: A focus group study with people with multiple sclerosis, carers and occupational therapists. J Health Psychol. 2019; 24: 480–494. https://doi.org/10.1177/1359105316677293 PMID: 27852887

29. Johnson KL, Bamer AM, Yorkston KM, Amtmann D. Use of cognitive aids and other assistive technology by individuals with multiple sclerosis. Disabil Rehabil Assist Technol. 2009; 4: 1–8. https://doi.org/10.1080/17483100802239648 PMID: 19172475

30. Lavelle G, Norris M, Flemming J, Harper J, Bradley J, Johnston H, et al. Validity and Acceptability of Wearable Devices for Monitoring Step-Count and Activity Minutes Among People With Multiple Sclerosis. Front Rehabil Sci. 2022; 2: 96. https://doi.org/10.3389/fresc.2021.737384 PMID: 36188762

31. Block VJ, Zhao C, Hollenbach JA, Olgin JE, Marcus GM, Pletcher MJ, et al. Validation of a consumer-grade activity monitor for continuous daily activity monitoring in individuals with multiple sclerosis. Mult Scler J—Exp Transl Clin. 2019; 5: 205521731988866. https://doi.org/10.1177/2055217319888660 PMID: 31803492

32. Johnston W, Judice PB, Molina García P, Mühlen JM, Lykke Skovgaard E, Stang J, et al. Recommendations for determining the validity of consumer wearable and smartphone step count: Expert statement and checklist of the INTERLIVE network. British Journal of Sports Medicine. BMJ Publishing Group; 2021. pp. 780–793. https://doi.org/10.1136/bjsports-2020-103147 PMID: 33361276

33. Feito Y, Garner HR, Bassett DR. Evaluation of ActiGraph's low-frequency filter in laboratory and free-living environments. Med Sci Sports Exerc. 2015; 47: 211–217. https://doi.org/10.1249/MSS.0000000000000395 PMID: 24870583

34. Webber SC, St John PD. Comparison of ActiGraph GT3X+ and Step watch step count accuracy in geriatric rehabilitation patients. J Aging Phys Act. 2016; 24: 451–458. https://doi.org/10.1123/japa.2015-0234 PMID: 26751505

35. Kozey-Keadle S, Libertine A, Lyden K, Staudenmayer J, Freedson PS. Validation of wearable monitors for assessing sedentary behavior. Medicine and Science in Sports and Exercise. Med Sci Sports Exerc; 2011. pp. 1561–1567. https://doi.org/10.1249/MSS.0b013e31820ce174 PMID: 21233777

36. Motl RW, Sasaki JE, Cederberg KL, Jeng B. Social-cognitive theory variables as correlates of sedentary behavior in multiple sclerosis: Preliminary evidence. Disabil Health J. 2019; 12: 622–627. https://doi.org/10.1016/j.dhjo.2019.05.002 PMID: 31130491

37. Neal WN, Cederberg KL, Jeng B, Sasaki JE, Motl RW. Is Symptomatic Fatigue Associated With Physical Activity and Sedentary Behaviors Among Persons With Multiple Sclerosis? Neurorehabil Neural Repair. 2020; 34: 505–511. https://doi.org/10.1177/1545968320916159 PMID: 32340521

38. Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. J Sci Med Sport. 2011; 14: 411–416. https://doi.org/10.1016/j.jsams.2011.04.003 PMID: 21616714

39. Carr LJ, Mahar MT. Accuracy of intensity and inclinometer output of three activity monitors for identification of sedentary behavior and light-intensity activity. J Obes. 2012; 2012: 1–9. https://doi.org/10.1155/2012/460271 PMID: 22175006

40. Blikman LJM, Van Meeteren J, Rizopoulos D, De Groot V, Beckerman H, Stam HJ, et al. Physical behaviour is weakly associated with physical fatigue in persons with multiple sclerosis-related fatigue. J Rehabil Med. 2018; 50: 821–827. https://doi.org/10.2340/16501977-2375 PMID: 30183054

41. Blikman LJ, Van Meeteren J, Horemans HL, Kortenhorst IC, Beckerman H, Stam HJ, et al. Is physical behavior affected in fatigued persons with multiple sclerosis? Arch Phys Med Rehabil. 2015; 96: 24–29. https://doi.org/10.1016/j.apmr.2014.08.023 PMID: 25239283

42. Activity Time Series. [cited 1 Nov 2022]. Available: https://dev.fitbit.com/build/reference/web-api/activity-timeseries/

43. Fitbit LLC. How does my Fitbit device calculate my daily activity? 2022 [cited 1 Nov 2022]. Available: https://help.fitbit.com/articles/en_US/Help_article/1141.htm

44. Sandroff BM, Riskin BJ, Agiovlasitis S, Motl RW. Accelerometer cut-points derived during over-ground walking in persons with mild, moderate, and severe multiple sclerosis. J Neurol Sci. 2014; 340: 50–57. https://doi.org/10.1016/j.jns.2014.02.024 PMID: 24635890

45. Eldemir K, Guclu-Gunduz A, Ozkul C, Eldemir S, Soke F, Irkec C. Associations between fatigue and physical behavior in patients with multiple sclerosis with no or minimal disability. Fatigue Biomed Heal Behav. 2021; 9: 69–78. https://doi.org/10.1080/21641846.2021.1923995

46. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955; 52: 281–302. https://doi.org/10.1037/h0040957 PMID: 13245896

47. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol. 1995; 50: 741–749. https://doi.org/10.1037/0003-066X.50.9.741

48. Cohen RJ, Swerdlik ME, Sturman ED. Psychological testing and assessment: an introduction to tests and measurement. 8th ed. New York, New York, USA: McGraw-Hill Professional; 2012.

49. Hattie J, Cooksey RW. Procedures for Assessing the Validities of Tests Using the "Known-Groups" Method. Appl Psychol Meas. 1984; 8: 295–305. https://doi.org/10.1177/014662168400800306

50. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychol Bull. 1959; 56: 81–105. https://doi.org/10.1037/h0046016 PMID: 13634291

51. von Wyl V, Gonzenbach R. Barriers to Physical Activity in People With MS (BarkA-MS). 2021 [cited 1 Apr 2022]. Available: https://clinicaltrials.gov/ct2/show/study/NCT04746807

52. Butland RJA, Pang J, Gross ER, Woodcock AA, Geddes DM. Two-, six-, and 12-minute walking tests in respiratory disease. Br Med J. 1982; 284: 1607–1608. https://doi.org/10.1136/bmj.284.6329.1607 PMID: 6805625

53. Casey B, Coote S, Donnelly A. Objective physical activity measurement in people with multiple sclerosis: a review of the literature. Disabil Rehabil Assist Technol. 2018; 13: 124–131. https://doi.org/10.1080/17483107.2017.1297859 PMID: 28285547

54. Sasaki JE, Sandroff B, Bamman M, Motl RW. Motion sensors in multiple sclerosis: Narrative review and update of applications. Expert Review of Medical Devices. Taylor and Francis Ltd; 2017. pp. 891–900. https://doi.org/10.1080/17434440.2017.1386550 PMID: 28956457

55. Sandroff BM, Motl RW, Suh Y. Accelerometer output and its association with energy expenditure in persons with multiple sclerosis. J Rehabil Res Dev. 2012; 49: 467–476. https://doi.org/10.1682/jrrd.2011.03.0063 PMID: 22773205

56. Polhemus AM, Haag C, Sieber C, Sylvester R, Kool J, Gozenbach R, et al. Methodological heterogeneity induces bias on physical activity metrics derived from the Actigraph GT3X in multiple sclerosis. Front Rehabil Sci. 2022;accepted.

57. Chikere CMU, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard—An update. PLoS One. 2019; 14: e0223832. https://doi.org/10.1371/journal.pone.0223832 PMID: 31603953

58. Williamson GR. Illustrating triangulation in mixed-methods nursing research. Nurse researcher. Nurse Res; 2005. pp. 7–18. https://doi.org/10.7748/nr2005.04.12.4.7.c5955 PMID: 16045043

59. Sandroff BM, Motl RW, Pilutti LA, Learmonth YC, Ensari I, Dlugonski D, et al. Accuracy of StepWatchTM and ActiGraph accelerometers for measuring steps taken among persons with multiple sclerosis. Paul F, editor. PLoS One. 2014; 9: e93511. https://doi.org/10.1371/journal.pone.0093511 PMID: 24714028

60. McGuigan C, Hutchinson M. Confirming the validity and responsiveness of the Multiple Sclerosis Walking Scale-12 (MSWS-12). Neurology. 2004; 62: 2103–2105. https://doi.org/10.1212/01.wnl.0000127604.84575.0d PMID: 15184625

61. Goldman MD, Ward MD, Motl RW, Jones DE, Pula JH, Cadavid D. Identification and validation of clinically meaningful benchmarks in the 12-item Multiple Sclerosis Walking Scale. Mult Scler. 2017; 23: 1405–1414. https://doi.org/10.1177/1352458516680749 PMID: 27903937

62. Hur SA, Guler SA, Khalil N, Camp PG, Guenette JA, Swigris JJ, et al. Minimal Important Difference for Physical Activity and Validity of the International Physical Activity Questionnaire in Interstitial Lung Disease. Ann Am Thorac Soc. 2019; 16: 107–115. https://doi.org/10.1513/AnnalsATS.201804-265OC PMID: 30211616

63. Kurtzke JF. Rating neurologic impairment in multiple sclerosis: An expanded disability status scale (EDSS). Neurology. 1983; 33: 1444–1452. https://doi.org/10.1212/wnl.33.11.1444 PMID: 6685237

64. Kempen J, De Groot V, Knol DL, Polman CH, Lankhorst GJ, Beckerman H. Community walking can be assessed using a 10-metre timed walk test. Mult Scler J. 2011; 17: 980–990. https://doi.org/10.1177/1352458511403641 PMID: 21622593

65. Goldman MD, Marrie RA, Cohen JA. Evaluation of the six-minute walk in multiple sclerosis subjects and healthy controls. Mult Scler. 2008; 14: 383–390. https://doi.org/10.1177/1352458507082607 PMID: 17942508

66. Motl RW, Sasaki JE, Cederberg KL, Jeng B. Validity of Sitting Time Scores From the International Physical Activity Questionnaire-Short Form in Multiple Sclerosis. Rehabil Psychol. 2019; 64: 463. https://doi.org/10.1037/rep0000280 PMID: 31107044

67. Sandroff BM, Motl RW. Comparison of ActiGraph activity monitors in persons with multiple sclerosis and controls. Disabil Rehabil. 2013; 35: 725–731. https://doi.org/10.3109/09638288.2012.707745 PMID: 23557239

68. Motl RW, Zhu W, Park Y, McAuley E, Scott JA, Snook EM. Reliability of scores from physical activity monitors in adults with multiple sclerosis. Adapt Phys Act Q. 2007; 24: 245–253. https://doi.org/10.1123/apaq.24.3.245 PMID: 17916920

69. Sandroff BM, Motl RW. Device-Measured Physical Activity and Cognitive Processing Speed Impairment in a Large Sample of Persons with Multiple Sclerosis. J Int Neuropsychol Soc. 2020; 26: 798–805. https://doi.org/10.1017/S1355617720000284 PMID: 32209162

70. Vanbelle S. Comparing dependent kappa coefficients obtained on multilevel data. Biometrical J. 2017; 59: 1016–1034. https://doi.org/10.1002/bimj.201600093 PMID: 28464322

71. Lin LI-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. 1989; 45: 255. https://doi.org/10.2307/2532051 PMID: 2720055

72. Altman DG, Bland JM. Measurement in Medicine: The Analysis of Method Comparison Studies. Stat. 1983; 32: 307. https://doi.org/10.2307/2987937

73. Parker RA, Scott C, Inácio V, Stevens NT. Using multiple agreement methods for continuous repeated measures data: A tutorial for practitioners. BMC Med Res Methodol. 2020; 20: 1–18. https://doi.org/10.1186/s12874-020-01022-x PMID: 32532218

74. Rosenthal R. Parametric measures of effect size. In: Cooper H, Hedges L, editors. The handbook of research synthesis. Russel Sage Foundation; 1994. pp. 231–244.

75. Altman D. Practical statistics for medical research. London, UK: Chapman and Hall; 1991.

76. Fleiss JL, Cohen J. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. Educ Psychol Meas. 1973; 33: 613–619. https://doi.org/10.1177/001316447303300309

77. Vetrovsky T, Siranec M, Marencakova J, Tufano JJ, Capek V, Bunc V, et al. Validity of six consumer-level activity monitors for measuring steps in patients with chronic heart failure. PLoS One. 2019;14. https://doi.org/10.1371/journal.pone.0222569 PMID: 31518367

78. Tedesco S, Sica M, Ancillao A, Timmons S, Barton J, O'Flynn B. Validity evaluation of the fitbit charge2 and the garmin vivosmart HR+ in free-living environments in an older adult cohort. JMIR mHealth uHealth. 2019;7. https://doi.org/10.2196/13084 PMID: 31219048

79. DeShaw KJ, Ellingson L, Bai Y, Lansing J, Perez M, Welk G. Methods for Activity Monitor Validation Studies: An Example With the Fitbit Charge. J Meas Phys Behav. 2018; 1: 130–135. https://doi.org/10.1123/jmpb.2018-0017

80. Reid RER, Insogna JA, Carver TE, Comptour AM, Bewski NA, Sciortino C, et al. Validity and reliability of Fitbit activity monitors compared to ActiGraph GT3X+ with female adults in a free-living environment. J Sci Med Sport. 2017; 20: 578–582. https://doi.org/10.1016/j.jsams.2016.10.015 PMID: 27887786

81. Knox KB, Clay L, Stuart-Kobitz K, Nickel D. Perspectives on walking from people with multiple sclerosis and reactions to video self-observation. Disabil Rehabil. 2018; 1–8. https://doi.org/10.1080/09638288.2018.1496154 PMID: 30348030

82. Raman S. Lived experience of walking in people with multiple sclerosis. University of Illinois at Chicago. 2010.

83. World Health Organization. Global Recommendations on Physical Activity for Health. 2009 [cited 22 Apr 2022]. Available: https://www.who.int/health-topics/physical-activity

84. Prince SA, Reid RD, Bernick J, Clarke AE, Reed JL. Single versus multi-item self-assessment of sedentary behaviour: A comparison with objectively measured sedentary time in nurses. J Sci Med Sport. 2018; 21: 925–929. https://doi.org/10.1016/j.jsams.2018.01.018 PMID: 29500119

85. Healy GN, Clark BK, Winkler EAH, Gardiner PA, Brown WJ, Matthews CE. Measurement of adults' sedentary time in population-based studies. American Journal of Preventive Medicine. Elsevier; 2011. pp. 216–227. https://doi.org/10.1016/j.amepre.2011.05.005 PMID: 21767730

86. Harada ND, Chiu V, King AC, Stewart AL. An evaluation of three self-report physical activity instruments for older adults. Med Sci Sports Exerc. 2001; 33: 962–970. https://doi.org/10.1097/00005768-200106000-00016 PMID: 11404662

87. Heesch KC, van Uffelen JGZ, Hill RL, Brown WJ. What do IPAQ questions mean to older adults? Lessons from cognitive interviews. Int J Behav Nutr Phys Act. 2010; 7: 35. https://doi.org/10.1186/1479-5868-7-35 PMID: 20459758

88. Kossoff Meredith, Ritter Jesse, Keller Jennifer, Ellen Mowry KM. P0888—Multiple Sclerosis Patients' Perceptions of Using an Accelerometer and Mobile App for Clinical Research. MS Virtual 2020. 2020.

89. Giunti G, Mylonopoulou V, Romero OR. More stamina, a gamified mHealth solution for persons with multiple sclerosis: Research through design. JMIR mHealth uHealth. 2018; 6: e9437. https://doi.org/10.2196/mhealth.9437 PMID: 29500159

90. Giunti G, Kool J, Rivera Romero O, Dorronzoro Zubiete E, Anderson K, Burford O, et al. Mobile Health Apps to Facilitate Self-Care: A Qualitative Study of User Experiences. PLoS One. 2016; 11: e0156164. https://doi.org/10.1371/journal.pone.0156164 PMID: 27214203

91. Polhemus AM, Novak J, Ferrão J, Simblett S, Radaelli M, Locatelli P, et al. Human-Centered Design Strategies for Device Selection in mHealth Programs: Development of a Novel Framework and Case Study. JMIR mHealth uHealth. 2020;8. https://doi.org/10.2196/16043 PMID: 32379055

92. Polhemus AM, Novak J, Majid S, Simblett S, Morris D, Bruce S, et al. Data Visualization for Chronic Neurological and Mental Health Condition Self-management: Systematic Review of User Perspectives. JMIR Ment Heal. 2022; 9: e25249. https://doi.org/10.2196/25249 PMID: 35482368

93. Fitbit Inspire 3 | Health & fitness tracker. [cited 2 Nov 2022]. Available: https://www.fitbit.com/global/us/products/trackers/inspire3?istCompanyId=a7a58ef0-2b29-4347-933b-7dd692310664&istFeedId=a6a412df-2601-466f-bd8f-5ece2201d669&istItemId=rxmtaxrrl&istBid=t&utm_medium=shopping&utm_source=google&utm_campaign=US_PF_ROAS&gclid=CjwKCAjwh4ObBhAzEiwAHzZYUyyGH-sS9hSfv7E9sKXBrZRZVI7Gv4TLjIMkjNlweWxAb7ShwMkEfxoCwSwQAvD_BwE&gclsrc=aw.ds