Brief Communication

Ger J Exerc Sport Res
https://doi.org/10.1007/s12662-022-00813-2
Received: 26 November 2021
Accepted: 7 April 2022

Kristof Van Laerhoven[1] · **Alexander Hoelzemann**[1] · **Iris Pahmeier**[2] · **Andrea Teti**[2] · **Lars Gabrys**[3]

[1] University of Siegen, Siegen, Germany
[2] ESAB Potsdam, Potsdam, Germany
[3] University of Vechta, Vechta, Germany

Validation of an open-source ambulatory assessment system in support of replicable activity studies

Introduction

Wearable devices that track physical activity and vital signs like heart rate for health promotion and monitoring have become prevalent, with fitness trackers or smartwatches enjoying a rise in popularity. In a recent Gallup poll, 19% of US adults report at some point having worn a fitness tracker or smartwatch (34%) or having tracked their health statistics on a phone or tablet (32%) (McCarthy 2019). Commercially available activity trackers are more and more used in clinical trials to assess the physical activity behavior of study participants over time. When the tracker's original sensor data are not recorded but instead preprocessed descriptors such as counts, steps, sedentary bouts, or activity levels are logged in the study, replication problems arise. Similarly, when such tracking devices are upgraded or have reached the end of their shelf-life, the hardware and algorithms that produced the data are often lost as well.

Consent to participate
Study participation was voluntary and informed consent was obtained from all participants.

Consent to publication
All participants gave written consent for publication.

Code availability
Our application source code for the Bangle.js smartwatch, our algorithms and the mobile app can be inspected and downloaded at: https://github.com/kristofvl/BangleApps and https://github.com/kristofvl/Activate2.

The past decades have seen a shift in the use of activity trackers for various studies and clinical trials. Objectively measured physical activity data are more accurate compared to subjective physical activity assessments like physical activity questionnaires (Garriguet et al 2015) with frequent overreporting of vigorous activities and underreporting of sedentary behavior (Fiedler et al 2021; Lines et al 2020; Verhoog et al 2019). For physical activity and health promotion strategies and interventions, the accurate recording and appropriate feedback to the user seem essential for intervention success and data interpretation. Among the commercially available trackers, the ActiGraph[1] series of inertial-based trackers are recognized as well validated and reliable devices and have thus far been predominantly used for a majority of published research studies (Wijndaele et al 2015).

Although the internal inertial sensors meanwhile come as standard chipset packages with well-understood data, are well-calibrated, and have common interfaces, commercial inertial-based trackers are commonly sold as black boxes, about little is known about the accuracy of their algorithms (such as step detectors) or inner workings (Brondin et al 2020). Studies have shown that there remain uncertainties about the reliability of the step counts reported by trackers, for instance, during low-intensity activities and when walking with assistive tools (Alinia et al 2017). Few algorithms that perform step detection in commercial inertial-based trackers' data are published as open-source code so that they can be reproduced, the work of Brondin et al (2020) being an exception.

With the advent of inertial-based trackers, algorithms have been developed to characterize the activity of the wearer over time. Whereas earlier studies focused on device-specific measures, such as counts from legacy devices in which a device-specific acceleration threshold is exceeded (Brønd et al 2017), more recent algorithms are increasingly defined on standard units that remain valid on any recordings of raw inertial sensor signals, even when the trackers are exchanged. In a study by Migueles et al (2019) on young adults wearing the ActiGraph GT3X inertial sensor, for instance, cut points were empirically found for dominant and nondominant wrist placement, based on the Euclidean norm minus one *g* (ENMO) (van Hees et al 2013). These values were found to be: sedentary time (less than 50 milli-*g*), light physical activity (50-–110 milli-*g*), moderate physical activity (110–440 milli-*g*), and vigorous physical activity (more than 440 milli-*g*). Other notable procedures include

[1] ActiGraph, Pensacola, FL: https://www.theactigraph.com/.

Published online: 05 May 2022

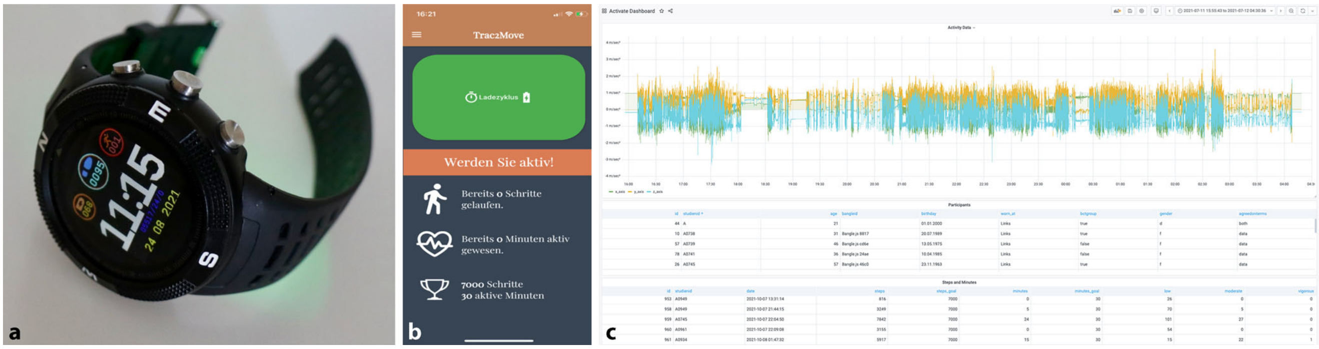

Fig. 1 ▲ The proposed components: data is collected and forwarded by an open-source smartwatch (Bangle.js), with our customized firmware **(a)** and smartphone app **(b)**, providing a transparent data pipeline to a back-end server database for anonymized storage and inspection of all data across participants **(c)**

mean amplitude deviation (Vähä-Ypyä et al 2015), High-pass filtered Euclidean norm (van Hees et al 2013), high-pass filtered Euclidean norm plus (van Hees et al 2013), proportional-integrating mode (PIM) (Jean-Louis et al 2001), zero crossing mode (Acebo et al 1999), and time above threshold (Fekedulegn et al 2020).

Similar efforts haven been under way for directly extracting other measures, such as accumulated number of steps, directly from the raw and normalized acceleration signals so that these become independent from the used sensors or hardware.

The aim of the studies presented is to further develop and evaluate an open-source ambulatory assessment system that produces data that can be complemented and compared against later. This consists of generic software and open algorithms for capturing physical activity under free living conditions, with a special focus on persons with noncommunicable diseases (diabetes, cardiovascular diseases), which can be used for research questions as well as health promotion strategies.

Methods

Our proposed system consists of an open-source smartwatch, which records raw accelerometer data simultaneously with locally computed aggregation measures as an immediate feedback to the wearer, a smartphone app that collects and forwards the collected data from the smartwatch at regular intervals, and a back-end server database. For all three components, we contribute with open-source software and algorithms.

Overview of components

The following sections will describe the specifications of the three main ambulatory assessment system components and motivate our choice for their selection.

Smartwatch. In order to decouple the recorded data from the specific devices that the data was recorded with, a fully open-source hardware and software design was used. The open source Bangle.js (■ Fig. 1a) is a smartwatch that is equipped with inertial sensors, a heart rate photoplethysmography (PPG) sensor, GPS, and a temperature sensor. It also sports a touch display and a vibration motor. Its processor is a Nordic 64 MHz nRF52832 ARM Cortex-M4 processor with Bluetooth LE, with 64 kB RAM 512 kB on-chip flash and 4 MB external flash. Both hardware design and embedded software (firmware) are publicly detailed, making it a fully replicable and customizable device, with all details of, for instance, the accelerometer sensor (a KIONIX KX023-1025[2]) known. The smartwatch's affordable cost of £ 54 in bulk and its robust IP-68 design furthermore allows for large-scale study deployments.

Smartphone app. We developed a custom smartphone app (■ Fig. 1b) as a cross-platform development for the current major mobile operating systems, iOS and Google Android, through implementation using Flutter, a cross-platform open-source app development tool by Google. It is designed to be easy to read and operate, and to seamlessly communicate with the smartwatch using Bluetooth 5 (low-energy) advertisements and serial communication on a regular basis. The interface consists of three main views and is displayed in German. It was designed to encourage participants to perform more physical activities in their daily lives and simultaneously collect physical activity data to be analyzed at a later date. Users can set their personal goals for the day and inspect their performance through a graphical overview of the daily metrics such as the number of steps, active minutes, and minutes spent in three levels of intensity categories (light, moderate, vigorous).

Server infrastructure. The server communicates with the client via two channels: information about the study participants, such as gender or age, and the confirmation of the consent form are sent via SSL and basic authentication to a reverse proxy, which then sends the information to the database via localhost. This reverse proxy communicates via a REST-API with a Postgres SQL database, where this information is then stored in an anonymized form. The recorded activity data, as well aggregated measures such as daily steps and active minutes, are sent daily via SSH to the server and stored in compressed binary files. At the server, these physical activity logs can then be processed further at a later date.

[2] https://www.kionix.com/product/KX023-1025 (Last access: 25 November 2021).

Ger J Exerc Sport Res https://doi.org/10.1007/s12662-022-00813-2

K. Van Laerhoven · A. Hoelzemann · I. Pahmeier · A. Teti · L. Gabrys

Validation of an open-source ambulatory assessment system in support of replicable activity studies

Abstract

Purpose: Inertial-based trackers have become a common tool in data capture for ambulatory studies that aim at characterizing physical activity. Many systems that perform remote recording of accelerometer data use commercial trackers and black-box aggregation algorithms, often resulting in data that are locked into proprietary formats and metrics that make later replication or comparison difficult.
Methods: The primary purpose of this manuscript is to validate an open-source ambulatory assessment system that consists of hardware devices, algorithms, and software components of our approach. We report on two validation experiments, one lab-based treadmill study on a convenience sample of 16 volunteers and one 'in vivo' study with 28 volunteers suffering from diabetes or cardiovascular disease.
Results: A comparison between data from ActiGraph GT9X trackers and our proposed system reveals that the original inertial sensor signals at the wrist strongly correlate (Pearson correlation coefficients for raw inertial sensor signals of 0.97 in the controlled treadmill-walking setting) and that estimated steps from an open-source wrist-based detection approach correlate with the hip-worn ActiGraph output (average Pearson correlation coefficients of 0.81 for minute-wise comparisons of detected steps) in day-long ambulatory data.
Conclusion: Recording inertial sensor data in a standardized form and relying on open-source algorithms on these data form a promising methodology that ensures that datasets can be replicated or enriched long after the wearable trackers have been decommissioned.

Keywords

Physical activity · Ambulatory assessment · Accelerometer · Open source · Activity trackers

Algorithms for data compression and aggregation

The firmware on the Bangle.js smartwatch contains, apart from custom routines to display information to the user and handle user interaction, event-based processes to (1) collect data from the onboard sensors, (2) compress and store these data efficiently on the local flash storage in time-stamped units, and (3) allow the wearer's smartphone to connect and collect the aggregate measures (such as steps, sedentary, or active minutes during the day), as well as download all recorded sensor data on a daily basis. the remainder of this article will focus on step detection as an example for such aggregation algorithms and measures.

The collection and loss-less compressing of accelerometer data are both tackled on the Bangle.js wristwatch. For every new 2-byte sample delivered by the accelerometer, a byte-wise delta-compression stage is executed, so that mostly incremental data is stored, where the most significant bytes (MSBs) are updated less frequently during normal usage. Encoding starts when subsequent samples all contain the same MSBs, storing these only once at the beginning together with the least significant bytes.

Few step detection algorithms have been published as open source in such a way that they can be straightforwardly implemented on raw acceleration data from a wrist-worn activity tracker. We have based our detection algorithm on empirically validated work published by Salvi et al (2018), which originally assumed smartphone-based accelerometer data and Brondin et al (2020) which presents a modification toward wrist-worn accelerometer data. The latter reports accuracy performances between 77% for slow-walking activity and 98% for outdoor walks, and its C source code has been made publicly available online[3] under an MIT licence. We have added a more stringent component as introduced by Salvi et al. (2018) to reduced false positive occurrences when the wearer's hand is inadvertently subjected to an impact.

Step-detection algorithms from three-dimensional accelerometer data commonly start by filtering the data with a low-pass FIR-filter to cut off the noise from frequencies above 3 Hz and calculating the magnitude of the acceleration vector that represents each sample in the data. This ensures that rotations of the accelerometer sensor are not regarded, and the pure impacts are visible as positive peaks in this new signal. The magnitude *mag* at time t is generally calculated from the readings at that time for all acceleration axes x, y, and z, r_x, r_y, r_z, as follows:

$$mag(t) = \sqrt{\sum_{a \in \{x,y,z\}} r_a^2}. \tag{1}$$

The result of this step is that subsequent *mag* readings can then be analyzed for positive peaks by searching for local maxima p at time t. These peaks are then, in what Brondin et al (2020) call a scoring stage, amplified by using the relative distances to the *mag* values before and after the occurring peak and following a window of size $2N$ around the time where the peak $p(t)$ occurred:

$$mag(t) = \frac{1}{2N} \times \sum_{k=-N,k\neq t}^{N} \left(mag(t) - mag(t+k)\right). \tag{2}$$

Windowed maximum peaks are then marked as steps by selecting candidate steps when the following equation holds, with μ the mean of the past *mag* samples and σ the standard deviation:

$$mag(t) - \mu > \sigma \cdot th. \tag{3}$$

The above algorithm steps were implemented in our Bangle.js firmware and additionally in Python to carry out offline experiments on the recorded and original 3D accelerometer readings, as detailed in the following sections.

[3] https://github.com/Oxford-step-counter/C-Step-Counter (last access: 4 March 2022).

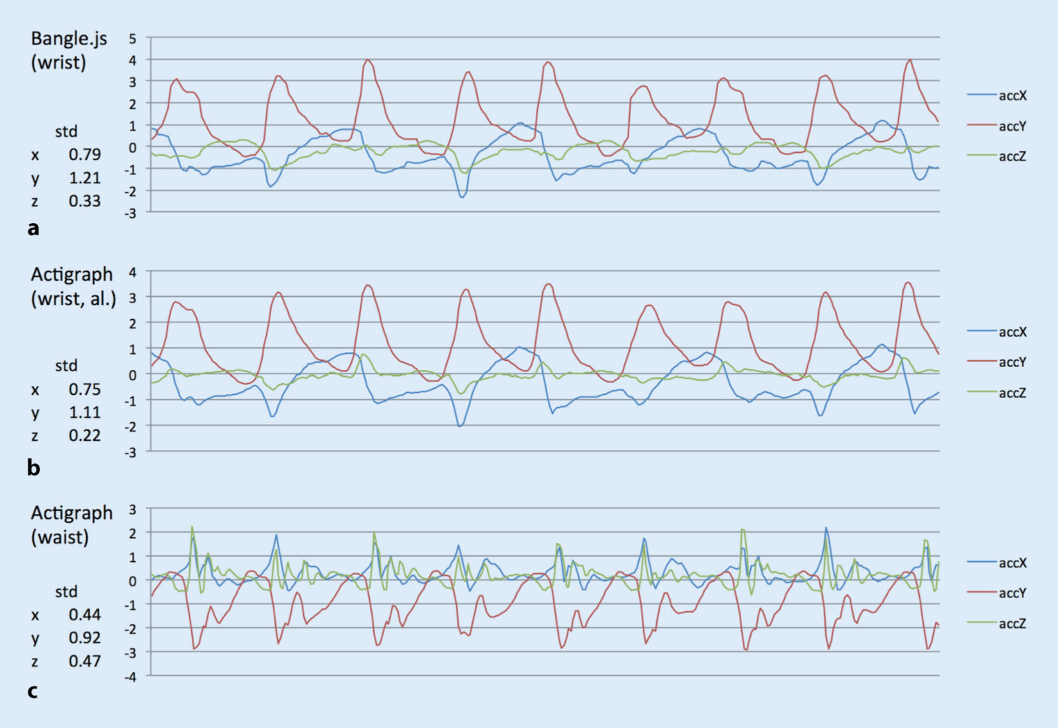

Fig. 2 ◀ Time series plots for a short segment of 3.4 s of accelerometer data (in g) from one participant's Bangle.js and ActiGraph GT9X devices, while running at vigorous speed (7 km h^{-1}) on the treadmill. The top (wrist) plots **(a, b)** show that after rotating the acceleration data, both devices produce very similar data; the Pearson correlation coefficient across all recording segments between the wrist-worn devices achieved 0.97. The presence of minor deviations in the amplitudes of the Y and Z axes' acceleration is likely due to slightly imperfect alignment between the two wrist-worn devices. The bottom plot **(c)** shows that acceleration at the hip shows larger differences in patterns

Experimental design

Two subsequent studies were performed to evaluate the use of our system: one laboratory study in which participants were asked to move on a treadmill at varying speeds to elicit different activity levels (light, moderate, vigorous) and to compare the raw accelerometer readings. A second study was subsequently performed to measure physical activity behavior for one day during participants' daily routines. In both studies, participants were asked to wear our customized Bangle.js tracker at the wrist and the ActiGraph GT9X sensor as the reference device, at the hip and at the wrist. The study was approved by the ethics committee of the medical association of Lower Saxony and was conducted in accordance with the declaration of Helsinki. Participation was always voluntary and informed consent obtained from all participants. To support the following data analysis and visualization routines, we developed and used a series of custom Python-based scripts that are supplied with this article's code. The Declarations section at the end of this manuscript provides details on where to obtain the source code.

Preliminary validation experiment (lab study)

The first study aims at comparing the raw inertial sensor data between the Bangle.js open-source wristwatch and the commonly deployed ActiGraph GT9X across different activity levels (light, moderate, vigorous). Both devices contain a 3D accelerometer, a 3D gyroscope, and a Bluetooth Low Energy (BLE) interface to obtain access to locally stored sensor data.

For this preliminary calibration study a convenient sample of 16 healthy volunteers (9w, 7m) with a mean age of 45.4 years (±6.8 SD) and a BMI of 24.2 kg m^{-2} (±3.1 SD) was recruited among university staff. Participants must be able to walk and run on a treadmill. Severe diseases or pain lead to study exclusion. All participants were asked to wear the Bangle.js and an ActiGraph GT9X device next to each other on the wrist of the nondominant hand, as well as an ActiGraph GT9X on the hip. These three devices were set to record raw acceleration data at a sampling rate of 100 Hz and at a sensitivity of ±8 g. Data was recorded on the three devices, converted to common-format comma separated values (CSV) files per experiment run, and synchronized in two phases: once using the local time stamps, and a subsequent fine-grained adjustment by hand using the single impact peaks occurring in all three devices' three-dimensional signals. To choose the activity intensity levels, we followed the Compendium of Physical Activities by Ainsworth et al (2011), to select three target speeds on a treadmill, each for 5 min: 3 km h^{-1} walking for light intensity, 5 km h^{-1} walking for moderate intensity, and 7 km h^{-1} jogging for vigorous intensity.

Daily routine physical activity experiment

With a focus on physical activity and health promotion for persons with noncommunicable diseases like cardiovascular diseases (CVD) or metabolic diseases like diabetes mellitus, the second study measured the daily amount of physical activity of 28 participants (14w, 14m) with diabetes or CVD under real-life conditions. Since the outcome of interest for this study was daily physical activity, and no additional activity was to be performed, no exclusion criteria were defined, for instance on the wear

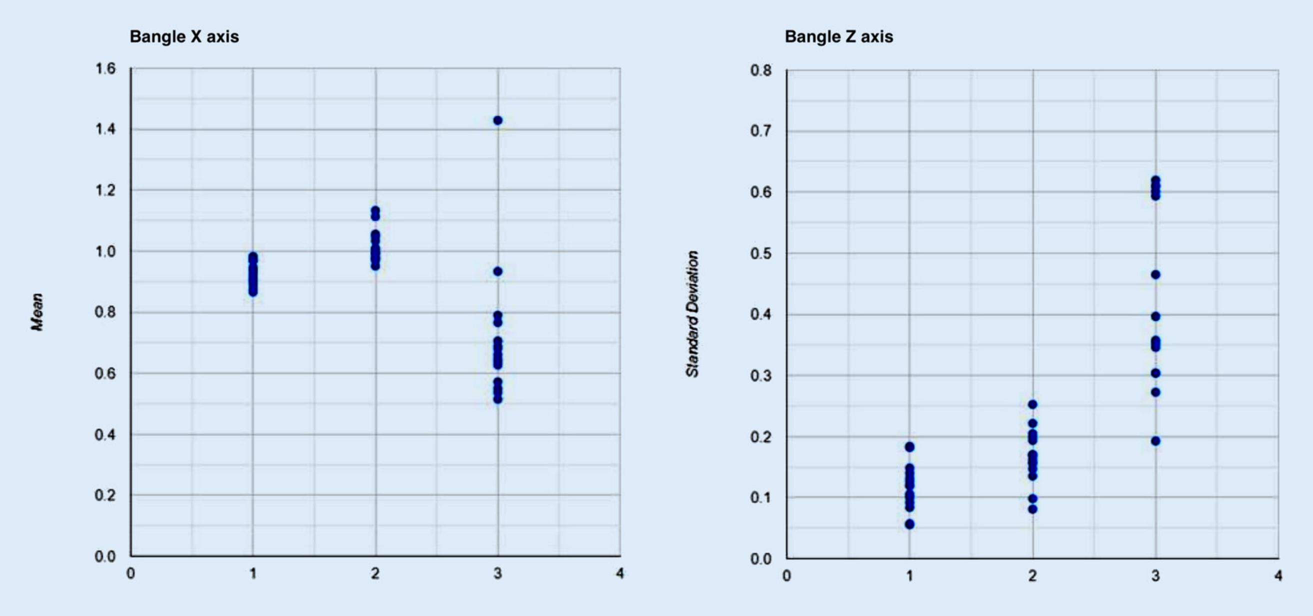

Fig. 3 ▲ Using the mean and standard deviation across sliding windows has shown that the mean over *X* and standard deviation over the *Z* axis values results in an optimal separation between low and moderate (accuracy: 97.5%), and moderate and vigorous (accuracy: 86.7%) activity intensities, respectively

time per day. Sample size approximation was done on the basis of previous studies with a comparable study design (Imboden et al 2018; Montoye et al 2020). Participants were recruited in a rehabilitation sports facility with a mean age of 65.6 years (±13.2 SD) and a BMI of 28.2 kg m^{-2} (±8.0 SD). Participants wore a Bangle.js at the wrist and the ActiGraph GT9X at their waist to record both raw accelerometer data. Using the ActiLife software framework, we post-processed the recordings from the ActiGraph in order to obtain ActiLife's aggregation metrics, such as steps, to compare our system's output to in terms of accuracy and deviations in values and time.

Experimental results

The following sections will present the outcome of the experiments as described in the previous section, using both visual inspection of the accelerometer signals, as well as through the aggregated performance metric of steps. All code for the analysis was written in Python and is made available online[4].

Preliminary validation experiment (lab study)

A first analysis of the wrist data between both devices shows that the alignment of acceleration axes requires that the y axes be identical but that the x axes and z axes be inverted. The two upper plots of ◘ **Fig. 2** show similar signals for all axes once the values of the ActiGraph are changed according to the following transformation: (x, y, z) to $(-x, y, -z)$, so that the internal three axes for both devices are aligned. A small difference in the signal amplitude for the y and z axes can be observed, which can be explained by the fact that the devices are not strapped in exactly the same orientation and position to the wearer's wrist. The time series are annotated on the left-hand side with the standard deviation values for single axes over larger windows, which are very similar for the wrist signals.

Sliding windowed mean and standard deviation statistics were used over all three accelerometer axes to obtain a set of optimal thresholds to separate between the three activity intensities. By following this procedure, the best-performing axes and features were found to be the mean over a window of minimally 3 s sliding over the X axis and the standard deviation over a window of minimally 4 s sliding over the Z axis. The classification accuracy obtained to distinguish between low-moderate or moderate-vigorous activity is 97.5 and 86.7%, respectively.

Using the ActiLife software on the recorded data from the ActiGraph GT9X, we obtained the resulting steps per participant per intensity class, for analysis and comparison of both the wrist-worn and the hip-worn devices' data. These steps data were obtained with time stamps for each second in the datasets, so that the resulting step counts could be compared between the hip-worn and the wrist-worn ActiGraph devices. After disregarding the data from transitional periods, for instance when participants moved from one speed to another on the treadmill or at the start and end of the exercises, the step counts between our open-source approach on the Bangle.js and the steps estimated by the ActiLife software for both the wrist and hip-worn ActiGraphs matched exactly. It is important to note here that in these laboratory-like settings, where steps occur within a controlled and vigorous activity, step detection is known to have become extremely reliable, as for instance noted in Bassett et al (2016). The following study will investigate such

[4] https://github.com/kristofvl/Activate2 (last access: 4 March 2022).

Table 1 Evaluation results for detected steps from the ActiLife software on the hip-worn ActiGraph data (ActiLife Steps), the open-source algorithm on the Bangle data using optimal parameters across participants (Bangle Steps), and for the participant-customized parameters (Bangle,C Steps). Parametrisation was done using the overall step counts as a target, hence the lower deviations in steps for the latter approach. Even though the overall correlation results remain similar, performance for single participants tended to fluctuate significantly

ID (day)	ID (person)	Duration (hours)	ActiLife Steps	Bangle Steps	Deviation (%)	Correlation	Bangle, C Steps	Deviation (%)	Correlation
000	001	8.86	2295	1914	16.60	0.72	2263	1.39	0.62
001	002	9.34	1608	3104	93.03	0.80	1664	3.48	0.93
002	002	8.86	2998	3496	16.61	0.79	3004	0.20	0.85
003	003	8.1	3771	3838	1.78	0.78	3838	1.78	0.78
004	004	8.93	11940	5337	55.30	0.53	12157	1.82	0.83
005	005	7.93	6387	5510	13.73	0.95	6582	3.05	0.92
006	005	8.06	4167	3676	11.78	0.94	4167	0.00	0.90
007	006	8.97	11718	4832	58.76	0.77	11237	4.10	0.74
008	007	10.03	3084	2582	16.28	0.80	3102	0.58	0.93
009	007	9.98	4283	3903	8.87	0.83	4299	0.37	0.86
010	008	2.21	302	302	0.00	0.55	302	0.00	0.55
011	009	9.53	5123	4588	10.44	0.87	5194	1.39	0.90
012	009	8.46	8597	6238	27.44	0.81	8614	0.20	0.81
013	010	7.08	3340	4494	34.55	0.73	3339	0.03	0.67
014	011	7.43	526	573	8.94	0.93	515	2.09	0.94
015	012	8.32	3407	4073	19.55	0.84	3578	5.02	0.72
016	013	8.92	4779	4060	15.04	0.84	4690	1.86	0.92
017	013	8.99	4638	3894	16.04	0.75	4616	0.47	0.88
019	015	6.33	725	1051	44.97	0.92	747	3.03	0.92
020	016	8.56	1248	1745	39.82	0.73	1252	0.32	0.58
021	017	8.77	2397	2066	13.81	0.89	2405	0.33	0.90
022	018	8.81	3608	3665	1.58	0.86	3665	1.58	0.86
023	019	7.99	1735	1800	3.75	0.80	1800	3.75	0.80
024	020	9.04	579	857	48.01	0.79	577	0.35	0.63
025	021	10.08	6944	7994	15.12	0.95	7014	1.01	0.94
026	021	9.72	5404	5603	3.68	0.98	5525	2.24	0.98
027	022	9.52	5453	4981	8.66	0.97	5384	1.27	0.97
028	022	9.3	3312	2620	20.89	0.72	3246	1.99	0.66
029	023	8.29	3988	3863	3.13	0.83	3979	0.23	0.69
030	024	5.8	3638	2990	17.81	0.74	3749	3.05	0.79
031	025	9.23	7920	6697	15.44	0.96	7952	0.40	0.92
	Mean	8.43	4190.77	3510.81	21.34	0.82	4076.75	1.53	0.82

results in less-structured daily-life conditions.

Daily routine physical activity experiment

The analysis of the daily routine data from the wrist-worn Bangle.js and hip-worn ActiGraph had to be prepared more meticulously, as both devices were programmed to start at a pre-defined time and date, and as the study participants were responsible for donning and doffing the sensors themselves. The ActiGraph was configured to record for 2 days continuously, whereas the Bangle.js was configured to start and stop the recordings each day at set times. Two participants also took off the hip-worn device in the middle of the recordings, which led to one of the data streams generating data that was only partially usable. An additional hurdle was that for some recordings, the ActiGraphs stopped recording after approximately 24 to 30 h. The resulting day-long recordings were then trimmed to those times where valid data existed between the two devices and where the study participants were demonstrably wearing both devices. The left-most columns in ▣ Table 1 show the resulting synchronized recordings and their duration in hours, as well as their total number of steps as delivered by the ActiLife software.

The synchronization of the data was also verified by manually observing similar impacts at the start and at the end of the recordings. Since both devices are equipped with a dedicated real-time clock, this was found to be very accurate already and did not need additional corrections within the data. Similarly, since both devices were set to record at the

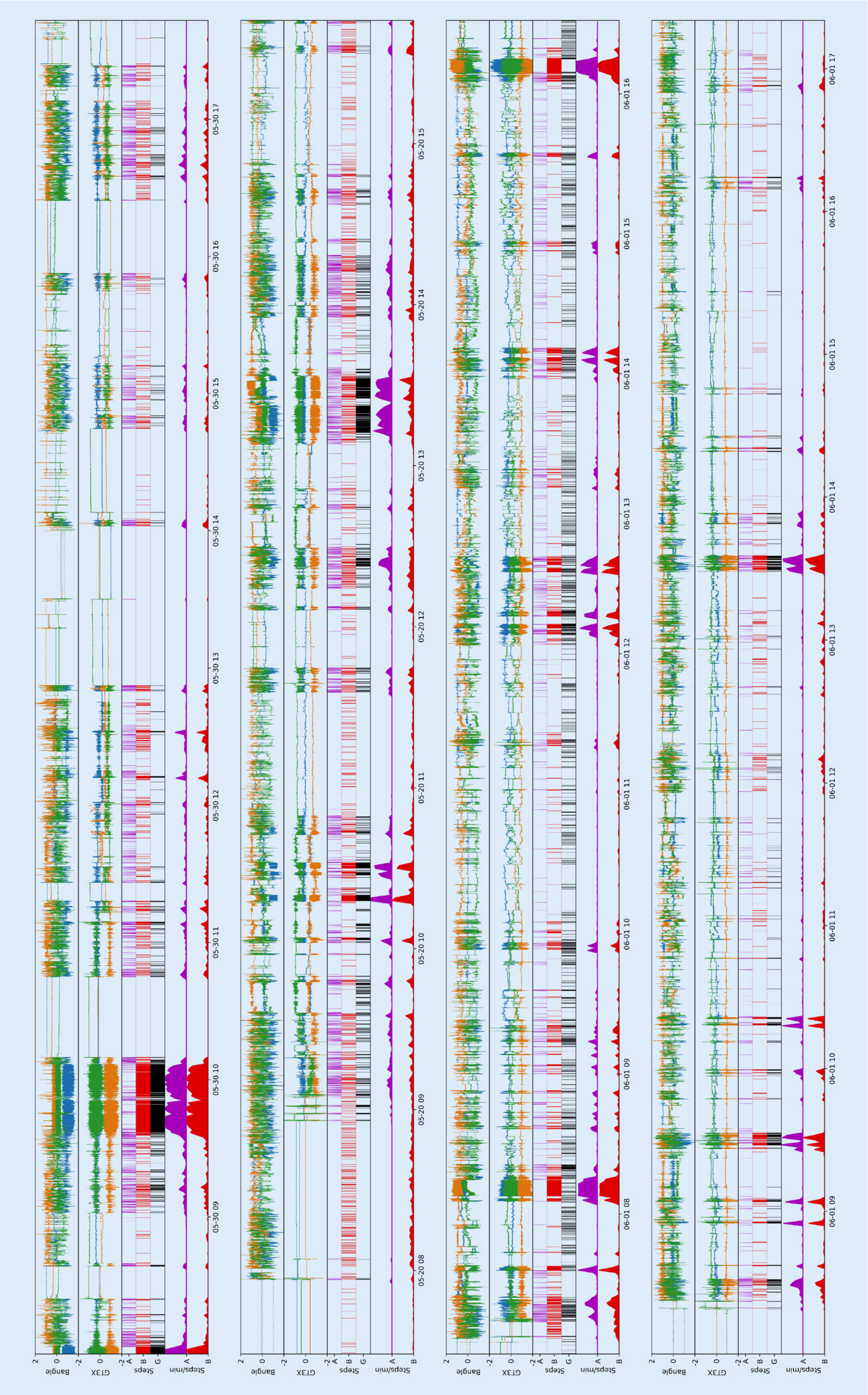

Fig. 4 ◀ Selection of well-performing recordings 026, 029, 027, and 002, displaying three-dimensional accelerometer data for the wrist-worn Bangle.js (top) and hip-worn ActiGraph (middle) devices (in g), with occurred and minute-accumulated steps (bottom). X labels: day, month, and hour of day

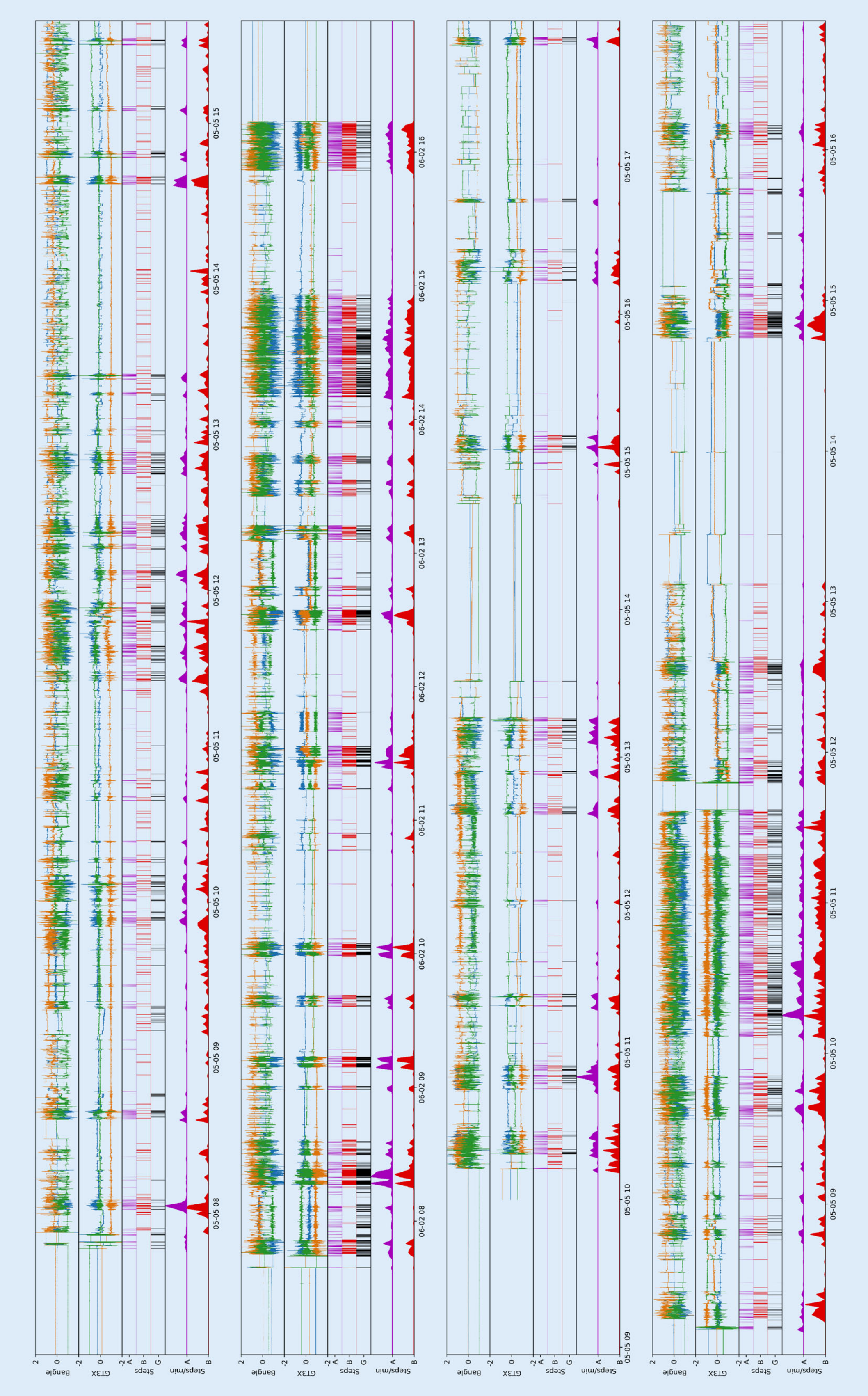

Fig. 5 ◀ Selection of poor-performing recordings 020, 010, 024, and 000, displaying three-dimensional accelerometer data for the wrist-worn Bangle.js (top) and hip-worn ActiGraph (middle) devices (in g), with occurred and minute-accumulated steps (bottom). X labels: day, month, and hour of day

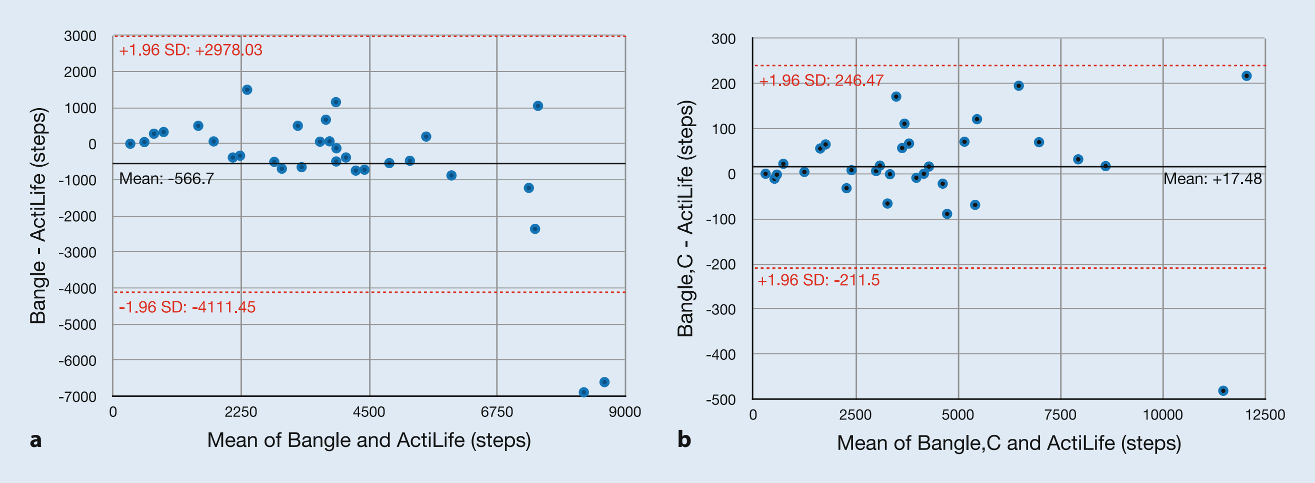

Fig. 6 ▲ Bland-Altman plots for visualizing differences in the steps as estimated by the ActiLife step detection algorithm on the hip-worn ActiGraph GT9X data, and the open-source algorithm on the wrist-worn Bangle.js data. We compare for the latter two cases: for the parameters chosen across all participants (Bangle in the left plot **a**), and for the parameters optimized per-participant (Bangle,C as depicted in the right plot **b**)

same ranges, the recorded values did not need any normalization.

Steps were detected in these experiments through three methods: (1) The hip-worn ActiGraph using the internal algorithm of the ActiLife software, to compare our values with a well-known method, (2) the detected steps as estimated by our open-source algorithm, using parameters optimally selected to match the number of steps of the ActiLife steps across all study participants, and (3) the steps as detected by our open-source algorithm, using parameters optimally selected to match the number of steps of the ActiLife steps per study participant. The distinction between the last two has been made to analyze the distinctive power of the algorithm's parameters.

The per-participant results can be viewed in **Table 1** in the middle and right-most columns. The steps detected by the ActiLife software on the ActiGraph device (hip-worn) or by the open-source algorithm as described above on the Bangle.js data were accumulated per minute in which they occurred. The Pearson correlation coefficient between the resulting minute-by-minute step count series was then calculated.

For the algorithm version where parameters were set in a participant-independent fashion, it can be seen that the minute-to-minute Pearson correlation coefficients tend to fluctuate significantly between both individual participants and monitored days. The Pearson correlation coefficient is above 0.9 for several day-segments, showing high correlation in steps predictions, and on average 0.818 when using the optimal parameters across all data segments or a slightly higher 0.819 for the optimal parameters selected per person. These results are in line with those of other studies that have compared commercial activity tracking devices, as for instance surveyed in Evenson et al (2015).

Figures 4 and 5 display the raw time series plots for selections of best-performing data recording segments, displaying the original three-dimensional accelerometer data for the wrist-worn Bangle.js (the top plot for each recording) and hip-worn ActiGraph (middle plot for each recording) devices, denoted in *g*. These are synchronized and combined with minute-accumulated steps in each bottom plot. The plots are annotated in the *X* axis with labels depicting the day, month, and hour of day. The *Y* axis is annotated with the source system of the step estimates: A for ActiLife-detected steps in the ActiGraph GT9X data, B for the steps detected by the presented open-source step detection algorithm on the Bangle.js data, and G for the steps detected by the presented open-source algorithm on the ActiGraph GT9X data.

The visualizations contain several examples where the steps detected from wrist-worn data tends to match well during longer bouts of walking, but shorter clusters of steps show more discrepancies. With the exception of the longer walking segment in recording 000 at the bottom in **Fig. 5**, from around 10 to 11:30, and the segment in recording 029 in **Fig. 4** from around 13:00 to 13:00, the recordings with poorer correlation coefficients between the ActiLife steps and those delivered by our own system tend to occur especially outside such events. Some of these occurrences can be tracked down to phases where the wrist data displays a lot of motion whereas the hip data shows significantly less motion, as for instance can be observed in **Fig. 4** for recording 002 at the bottom around 12:00 (with our Bangle.js-based steps displaying many steps, unlike the ActiGraph-based setup).

Figure 6 presents the Bland-Altman plots for further analyzing the agreement between the steps as found by the ActiLife algorithm within the hip-worn ActiGraph GT9X data, and the open source step detection algorithm as discussed in Sect. 2.2 on the wrist-worn Bangle data. As in **Table 1**, it can be seen that differences in steps between the two devices and algorithms are about an order of magnitude larger for the cross-participants parameters but also that even for those, larger step counts still show relatively small differences. Further examination using equivalence testing, as recently suggested in O'Brien (2021), would be a fitting methodology to demonstrate the equivalence between the ActiGraph-based hip-worn ActiLife steps detection and the Bangle.js-based wrist-worn approach presented here. As this requires additional data preparation and decisions, such as determining an equivalence interval, this is left for future research on our public dataset.

Conclusions and discussion

In summary, we proposed an approach that relies on strictly open-source components in order to be able to record study data that can be re-analyzed and supplemented afterwards, even after the used devices and software have reached their end-of-life limit or have gone out of production. For mere collection, this means that raw acceleration data needs to be recorded at known units of sensitivity and sampling, but for studies in which devices offer direct feedback to the participants, for instance displaying the steps taken during the day, this also requires open-source algorithms.

The purpose of the studies presented here was to validate our approach of using a low-cost smartwatch and compare the output, both the original sensor values and the aggregated measure such as steps. We took recommendations of the framework to evaluate devices that assess physical activity behavior into account, with a focus on phase I (laboratory development) and phase III (naturalistic validation) (Keadle et al 2019). Experimental results can be summarized into the following points. (1) The three-dimensional accelerometer data, when sampled at similar rates and set to the same range, are highly comparable between devices. The preliminary study where study volunteers walk and run at different speeds on a treadmill show an almost perfect correlation (Pearson correlation coefficient: 0.97) between the wrist-worn devices. (2) The detected steps in natural environments from the open-source algorithm from our custom wrist-worn tracker show strong correlation with the detected steps from a hip-worn ActiGraph using the ActiLife software suite. Deviations in step counts between the two systems can be seen by visual inspection especially in periods when study participants were only walking for brief, intermittent periods.

Our studies focused mostly on raw acceleration signals and evaluated step counts as measured by our algorithms and compared against the ActiLife software suite. Although the Actigraph GT9X is not the gold standard for validating steps, it is probably one of the most frequently used devices in physical activity research and, therefore, seems appropriate for study purposes. This is by far not the only measure that can be put to use for assessing physical activity in daily life. Future research will inadvertently have to focus on further measures that are both fast to comprehend by participants and can be calculated from any given acceleration signal. Robust estimates for specific exercises and activities, energy expenditure, or calorie consumption could be candidates for such measures and are similarly developed as an open-source code base.

Beyond the recording of activity data in a replicable way, systems such as the one presented in this manuscript would form an optimal basis for designing future just-in-time adaptive interventions, where events in the sensor data can trigger interventions on users' phones or smartwatches, as presented in Ebner-Priemer et al (2013) and Giurgiu et al (2020) for instance for smartphone-based triggers based on sedentary behavior and physical activity. Interventions based on events detected in the sensor data could be integrated directly on the smartwatch, thus guaranteeing that participants would immediately become notified, but would need additional attention to the usability and the reduced interaction modalities on such devices.

The anonymized experiment data described in this article will be placed online so that they are made available publicly for further analysis and comparisons: https:// github.com/kristofvl/Activate2

Corresponding address

Kristof Van Laerhoven
University of Siegen
57076 Siegen, Germany
kvl@eti.uni-siegen.de

Funding. This study is part of the project ActiVAtE_prevention (VW-ZN3426) which is funded by the Ministry for Science and Culture of the federal state of Lower Saxony.

Author Contribution. All authors contributed equally to this work.

Funding. Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest. The authors declare that they have no conflict of interests.

The study was approved by the medical chamber of the federal state of Lower Saxony.

References

Acebo, C., Sadeh, A., Seifer, R., Tzischinsky, O., Wolfson, A.R., Hafer, A., & Carskadon, M.A. (1999). Estimating sleep patterns with activity monitoring in children and adolescents: how many nights are necessary for reliable measures? *Sleep*, *22*(1), 95–103. https://doi.org/10.1093/sleep/22.1.95.

Ainsworth, B., Haskell, W., Herrmann, S., Meckes, N., Bassett, J.D., Tudor-Locke, C., Greer, J., Vezina, J., Whitt-Glover, M., & Leon, A. (2011). 2011 compendium of physical activities. *Medicine & Science in Sports & Exercise*, *43*(8), 1575–1581. https://doi.org/10.1249/mss.0b013e31821ece12.

Alinia, P., Cain, C., Fallahzadeh, R., Shahrokni, A., Cook, D., & Ghasemzadeh, H. (2017). How accurate is your activity tracker? a comparative study of step counts in low-intensity physical activities. *JMIR Mhealth Uhealth*, *5*(8), e106. https://doi.org/10.2196/mhealth.6321.

Bassett, D.R., Toth, L.P., LaMunion, S.R., & Crouter, S.E. (2016). Step counting: a review of measurement considerations and health-related applications. *Sports Medicine*, *47*(7), 1303–1315. https://doi.org/10.1007/s40279-016-0663-1.

Brønd, J. C., Andersen, L. B., & Arvidsson, D. (2017). Generating ActiGraph counts from raw acceleration recorded by an alternative monitor. *Medicine & Science in Sports & Exercise*, *49*(11), 2351–2360. https://doi.org/10.1249/mss.0000000000001344.

Brondin, A., Nordström, M., Olsson, C. M., & Salvi, D. (2020). Open source step counter algorithm for wearable devices. In *10th International Conference on the Internet of Things Companion*. IoT '20 Companion. New York: Association for Computing Machinery. https://doi.org/10.1145/3423423.3423431.

Ebner-Priemer, U., Koudela, S., Mutz, G., & Kanning, M. (2013). Interactive multimodal ambulatory monitoring to investigate the association between physical activity and affect. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2012.00596. https://www.frontiersin.org/article/10.3389/fpsyg.2012.00596.

Evenson, K. R., Goto, M. M., & Furberg, R. D. (2015). Systematic review of the validity and reliability of consumer-wearable activity trackers. *International Journal of Behavioral Nutrition and Physical Activity*. https://doi.org/10.1186/s12966-015-0314-1.

Fekedulegn, D., Andrew, M. E., Shi, M., Violanti, J. M., Knox, S., & Innes, K. E. (2020). Actigraphy-based assessment of sleep parameters. *Annals of Work Exposures and Health*, *64*(4), 350–367. https://doi.org/10.1093/annweh/wxaa007. https://doi.org/10.1093/annweh/wxaa007, https://academic.oup.com/annweh/article-pdf/64/4/350/33147831/wxaa007.pdf.

Fiedler, J., Eckert, T., Burchartz, A., Woll, A., & Wunsch, K. (2021). Comparison of self-reported and device-based measured physical activity using measures of stability, reliability, and validity in adults and children. *Sensors*. https://doi.org/10.3390/s21082672.

Garriguet, D., Tremblay, S., & Colley, R. C. (2015). Comparison of physical activity adult questionnaire results with accelerometer data. *Health reports*, *26*(7), 11–17.

Giurgiu, M., Plotnikoff, R. C., Nigg, C. R., Koch, E. D., Ebner-Priemer, U. W., & Reichert, M. (2020). Momentary mood predicts upcoming real-life sedentary behavior. *Scandinavian Journal of Medicine & Science in Sports*, *30*(7), 1276–1286. https://doi.org/10.1111/sms.13652.

van Hees, V. T., Gorzelniak, L., León, D. E. C., Eder, M., Pias, M., Taherian, S., Ekelund, U., Renström, F., Franks, P. W., Horsch, A., & Brage, S. (2013). Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PLOS ONE*, *8*(4), 1–10. https://doi.org/10.1371/journal.pone.0061691, https://doi.org/10.1371/journal.pone.0061691.

Imboden, M. T., Nelson, M. B., Kaminsky, L. A., & Montoye, A. H. (2018). Comparison of four fitbit and jawbone activity monitors with a research-grade actigraph accelerometer for estimating physical activity and energy expenditure. *British Journal of Sports Medicine*, *52*(13), 844–850.

Jean-Louis, G., Kripke, D. F., Mason, W. J., Elliott, J. A., & Youngstedt, S. D. (2001). Sleep estimation from wrist movement quantified by different actigraphic modalities. *Journal of Neuroscience Methods*, *105*(2), 185–191. https://doi.org/10.1016/s0165-0270(00)00364-2.

Keadle, S. K., Lyden, K. A., Strath, S. J., Staudenmayer, J. W., & Freedson, P. S. (2019). A framework to evaluate devices that assess physical behavior. *Exercise and sport sciences reviews*, *47*(4), 206–214.

Lines, R. L., Ntoumanis, N., Thøgersen-Ntoumani, C., McVeigh, J. A., Ducker, K. J., Fletcher, D., & Gucciardi, D. F. (2020). Cross-sectional and longitudinal comparisons of self-reported and device-assessed physical activity and sedentary behaviour. *Medicine & Science in Sports & Exercise*, *23*(9), 831–835. https://doi.org/10.1016/j.jsams.2020.03.004.

McCarthy, J. (2019). One in five u.s. adults use health apps, wearable trackers. Gallup Poll. https://news.gallup.com/poll/269096/one-five-adults-health-apps-wearable-trackers.aspx. Accessed: 2. May 2022

Migueles, J. H., Cadenas-Sanchez, C., Rowlands, A. V., Henriksson, P., Shiroma, E. J., Acosta, F. M., Rodriguez-Ayllon, M., Esteban-Cornejo, I., Plaza-Florido, A., Gil-Cosano, J. J., Ekelund, U., van Hees, V. T., & Ortega, F. B. (2019). Comparability of accelerometer signal aggregation metrics across placements and dominant wrist cut points for the assessment of physical activity in adults. *Scientific Reports*. https://doi.org/10.1038/s41598-019-54267-y.

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

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

Salvi, D., Velardo, C., Brynes, J., & Tarassenko, L. (2018). An optimised algorithm for accurate steps counting from smart-phone accelerometry. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 4423–4427). https://doi.org/10.1109/EMBC.2018.8513319.

Vähä-Ypyä, H., Vasankari, T., Husu, P., Mänttäri, A., Vuorimaa, T., Suni, J., & Sievänen, H. (2015). Validation of cut-points for evaluating the intensity of physical activity with accelerometry-based mean amplitude deviation (mad). *PLOS ONE*, *10*(8), 1–13. https://doi.org/10.1371/journal.pone.0134813.

Verhoog, S., Gubelmann, C., Guessous, I., Bano, A., Franco, O. H., & Marques-Vidal, P. (2019). Comparison of the physical activity frequency questionnaire (PAFQ) with accelerometry in a middle-aged and elderly population: the CoLaus study. *Maturitas*, *129*, 68–75. https://doi.org/10.1016/j.maturitas.2019.08.004.

Wijndaele, K., Westgate, K., Stephens, S. K., Blair, S. N., Bull, F. C., Chastin, S. F. M., Dunstan, D. W., Ekelund, U., Esliger, D. W., Freedson, P. S., Granat, M. H., Matthews, C. E., Owen, N., Rowlands, A. V., Sherar, L. B., Tremblay, M. S., Troiano, R. P., Brage, S., & Healy, G. N. (2015). Utilization and harmonization of adult accelerometry data. *Medicine and Science in Sports and Exercise*, *47*(10), 2129–2139. https://doi.org/10.1249/mss.0000000000000661.