

Recommendations for determining the validity of consumer wearable and smartphone step count: expert statement and checklist of the INTERLIVE network.

Title: Recommendations for determining the validity of consumer wearable and smartphone step count: expert statement and checklist of the INTERLIVE network.

Publication year: 2021

Journal: British journal of sports medicine

Authors: Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland.

Society: European Respiratory Society

Origin country: International (Europe)

Target countries: Europe

DOI: 10.1136/bjsports-2020-103147

PMID: 33361276

Source: <https://pubmed.ncbi.nlm.nih.gov/33361276/>

Br J Sports Med Br J Sports Med 694 bmjgroup 0432520 bjsm British Journal of Sports Medicine 0306-3674 1473-0480 pmc-is-collection-domain yes pmc-collection-title BMJ Open Access PMC8273687 PMC8273687.1 8273687 8273687 33361276 10.1136/bjsports-2020-103147 bjsports-2020-103147 1 Consensus Statement 1506 2314 Recommendations for determining the validity of consumer wearable and smartphone step count: expert statement and checklist of the INTERLIVE network <http://orcid.org/0000-0003-0525-6577> Johnston William 1 2 <http://orcid.org/0000-0003-2791-258X> Judice Pedro B 3 4 <http://orcid.org/0000-0001-6888-0997> Molina García Pablo 5 <http://orcid.org/0000-0001-9983-6403> Mühlen Jan M 6 Lykke Skovgaard Esben 7 Stang Julie 8 <http://orcid.org/0000-0001-9605-3489> Schumann Moritz 6 9 Cheng Shulin 6 9 Bloch Wilhelm 6 <http://orcid.org/0000-0001-6718-3022> Brønd Jan Christian 7 <http://orcid.org/0000-0003-2115-9267> Ekelund Ulf 8 <http://orcid.org/0000-0003-1584-679X> Grøntved Anders 7 Caulfield Brian 1 2 <http://orcid.org/0000-0003-2001-1121> Ortega Francisco B 5 <http://orcid.org/0000-0002-6230-6027> Sardinha Luis B 4 1 Insight Centre for Data Analytics , University College Dublin , Dublin , Ireland 2 School of Public Health, Physiotherapy and Sports Science , University College Dublin , Dublin , Ireland 3 Centro de Investigação em Desporto, Educação Física e Exercício e Saúde, CIDEFES , Universidade Lusófona , Lisbon , Portugal 4 Exercise and Health Laboratory, CIPER, Faculdade de Motricidade Humana , Universidade de Lisboa , Cruz-Quebrada , Portugal 5 PROFITH (PROmoting FITness and Health through physical activity) Research Group, Department of Physical Education and Sports, Faculty of Sport Sciences, Sport and Health University Research Institute (iMUDS) , University of Granada , Granada , Spain 6 Institute of Cardiovascular Research and Sports Medicine, Department of Molecular and Cellular Sports Medicine , German Sport University , Cologne , Germany 7 Department of Sports Science and Clinical Biomechanics, Research Unit for Exercise Epidemiology, Centre of Research in Childhood Health , University of Southern Denmark , Odense M , Denmark 8 Department of Sport Medicine , Norwegian School of Sport Sciences , Oslo , Norway 9 Exercise Translational Medicine Centre, the Key Laboratory of Systems Biomedicine, Ministry of Education, and Exercise, Health and Technology Centre, Department of Physical Education , Shanghai Jiao Tong University , Shanghai , China Correspondence to Professor Luis B Sardinha, Faculdade de Motricidade Humana, Universidade de Lisboa, Cruz-Quebrada 1499-002 Lisboa, Portugal; lsardinha@fmh.ulisboa.pt 7 2021 24 12 2020 55 14 385536 780 793 24 11 2020 12 07 2021 23 07 2021 24 07 2021 © Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY. Published by BMJ. 2021 <https://creativecommons.org/licenses/by/4.0/> This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of

whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/> . Consumer wearable and smartphone devices provide an accessible means to objectively measure physical activity (PA) through step counts. With the increasing proliferation of this technology, consumers, practitioners and researchers are interested in leveraging these devices as a means to track and facilitate PA behavioural change. However, while the acceptance of these devices is increasing, the validity of many consumer devices have not been rigorously and transparently evaluated. The Towards Intelligent Health and Well-Being Network of Physical Activity Assessment (INTERLIVE) is a joint European initiative of six universities and one industrial partner. The consortium was founded in 2019 and strives to develop best-practice recommendations for evaluating the validity of consumer wearables and smartphones. This expert statement presents a best-practice consumer wearable and smartphone step counter validation protocol. A two-step process was used to aggregate data and form a scientific foundation for the development of an optimal and feasible validation protocol: (1) a systematic literature review and (2) additional searches of the wider literature pertaining to factors that may introduce bias during the validation of these devices. The systematic literature review process identified 2897 potential articles, with 85 articles deemed eligible for the final dataset. From the synthesised data, we identified a set of six key domains to be considered during design and reporting of validation studies: target population, criterion measure, index measure, validation conditions, data processing and statistical analysis. Based on these six domains, a set of key variables of interest were identified and a 'basic' and 'advanced' multistage protocol for the validation of consumer wearable and smartphone step counters was developed. The INTERLIVE consortium recommends that the proposed protocol is used when considering the validation of any consumer wearable or smartphone step counter. Checklists have been provided to guide validation protocol development and reporting. The network also provide guidance for future research activities, highlighting the imminent need for the development of feasible alternative 'gold-standard' criterion measures for free-living validation. Adherence to these validation and reporting standards will help ensure methodological and reporting consistency, facilitating comparison between consumer devices. Ultimately, this will ensure that as these devices are integrated into standard medical care, consumers, practitioners, industry and researchers can use this technology safely and to its full potential.

Over the last decade, wearable and smartphone devices have become ubiquitous among us, with global wearable device and smartphone sales reaching an estimated 305.2 million and 1.5 billion worldwide in 2019, respectively. 1 2 These consumer devices afford individuals the ability to easily and objectively monitor their physical activity (PA) without the need for expensive laboratory grade equipment. 3 With the increasing proliferation of this technology, sports and exercise medicine practitioners and the general public are interested in leveraging these devices as a means to track and facilitate PA behavioural change. 4

One of the most common measures of PA levels provided by these consumer devices is step count. 5 Step count is easily understandable by the general population and there is a growing body of evidence supporting the use of these technological approaches in 'passive' PA monitoring and 'active' intervention. The 2018 PA Guidelines Advisory Committee Report highlighted that step counting is an accessible means to monitor and set PA goals, and that recent evidence supports an inverse dose–response relationship between daily step count and all-cause mortality, cardiovascular events and type 2 diabetes. 5–9 Furthermore, 'active' interventions using wearable activity trackers as a measurement tool can result in significant increases in PA participation, highlighting their potential utility in personalised medicine, increasing adherence to PA and to embed sustained healthy lifestyle habits. 4 10 Furthermore, the 'passive' and unobtrusive measurement of step count across the globe opens a new era of opportunities in the field of digital phenotyping. This population-level approach has recently been used to quantify regional inequalities in PA 11 and changes in PA levels due to public health containment measures introduced to combat a global pandemic. 12

While the increasing acceptance across the health and fitness industry has resulted in a surge in validation studies, 3 13–17 much of the published research fails to rigorously evaluate validity, and there is a lack of consistency across the published protocols, limiting valid comparisons between

devices. A letter to the British Journal of Sports Medicine called for 'evidence-based marketing claims', recommending that devices are thoroughly and transparently validated, to ensure that wearable technology can be used safely and to its full potential.¹⁸ In an attempt to develop a standardised validation process and certification, the Consumer Technology Association (CTA) developed a set of guidelines for the validation of wearable and/or app-based step counters.¹⁹ While a step in the right direction, the protocol proposed by the CTA focused solely on controlled treadmill walking and running in healthy individuals, failing to consider ecological validation of the technologies within free-living conditions and in cohorts with diverse gait characteristics. A major consequence of this is that while many of the studies demonstrate the validity of these devices during controlled trials, step counting algorithms typically have a higher margin of error during uncontrolled activities of daily living. For example, when considering the same wrist-worn wearable activity tracker compared with a gold-standard video criterion, a mean absolute percentage error (MAPE) of -0.7% – 5.2% was observed during controlled treadmill walking,²⁰ while a higher MAPE of 17.5% – 22.8% was seen during free-living conditions that are more representative of daily life.²¹

To address the above concerns, the Towards Intelligent Health and Well-Being Network of Physical Activity Assessment (INTERLIVE) was formed to develop a best-practice consumer wearable and smartphone step counter validation protocol. To do so, we conducted a systematic literature review of the methods that have previously been used to validate these consumer devices, identified and synthesised other relevant scientific literature to identify aspects that influence validation and developed best-practice validation recommendations. These recommendations are designed for device manufacturers, scientific institutions, healthcare providers and members of the general public who are interested in the transparent and rigorous validation of these devices.

INTERLIVE is a joint initiative of the University of Lisbon (Portugal), German Sport University (Germany), University of Southern Denmark (Denmark), Norwegian School of Sport Sciences (Norway), University College Dublin (Ireland), University of Granada (Spain) and Huawei Technologies, Finland. The consortium was founded in 2019 and strives towards developing best-practice methods for evaluating the validity of consumer wearables and smartphones. Moreover, we are aiming to increase awareness of the advantages and limitations of different validation methods and to introduce novel health-related metrics, fostering a widespread use of PA indicators.

An initial meeting was held in Cascais, Portugal on 15 November 2019, whereby the INTERLIVE consortium identified the development of best-practice validation protocols for consumer wearable and smartphone devices for step counting (part A) and heart rate monitoring (part B) as a key aim of the group. Within the consortium, it was agreed that the optimal process for developing the best-practice validation protocol should begin with extracting key elements of validation protocols previously used in the scientific literature. This information could then be used as the foundation for discussing the optimal and feasible protocols for conducting validation assessment that describes the accuracy end-users can expect. The consortium then formed two working groups: (1) heart rate monitoring (JMM, ELS, JS, SC, WB, JCB, UE, AG and MS) and (2) step counting (WJ, PBJ, PMG, BC, FBO and LBS). The working groups then developed the systematic literature review search strategies, prior to sharing them with the wider consortium. A second consortium meeting was held virtually on 10 March 2020 to finalise the search strategies, including the selection of the minimum a priori required criterion measure(s). The working groups then conducted the systematic literature search process and a framework was developed for extracting data, focusing on target population, criterion and index devices, testing conditions, data processing and statistical analysis. In parallel, additional searches of the wider literature were conducted to identify studies that highlighted factors that may affect the accuracy of wearable and smartphone based step counting that were outside of the scope of our search strategies. The data extraction process was then completed and multiple working group meetings were held to discuss each aspect of the validation protocols used within the included studies. A set of key domains for the best-practice recommendations were developed based on the synthesised outcomes of the literature review, the a priori knowledge relating to research grade device validation^{22–25} and the evidence-based expert opinion of the INTERLIVE members. The synthesised data were then reviewed with respect to these domains, and expert

validation protocols for wearable and smartphone step counters (part A) and wearable heart rate monitors (part B) were iteratively developed and shared with the entire consortium. At a virtual meeting held on 17 June 2020, the revised drafts were discussed and the protocols were aligned to ensure harmonisation of the statements. The revised drafts were then edited for consistency and reviewed by the wider consortium prior to circulation for final approval.

The primary aim of the initial systematic literature review search was to identify which methods and validation protocols are currently used in the scientific literature to validate consumer wearable and smartphone step counting devices. It was not within the scope of this review to examine the validity results of the included studies. The systematic literature review was conducted in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement and registered with the International Prospective Register of Systematic Reviews (ID: CRD42020177263). Specific search terms were used to identify peer-reviewed journal articles published in three electronic databases: PubMed, Embase and Web of Science. The search terms are presented in online supplemental appendix 1. To be eligible for inclusion, the studies were required to investigate the step count validity of a consumer wearable and/or smartphone, and compare it to a criterion measure such as video recording, visual step counting (ie, laboratory and semifree-living protocols) and/or any device or method including accelerometers, pedometers and wearable garments (ie, free-living protocols). Studies published from inception to 13 March, 2020, across any population, were deemed eligible for inclusion. Studies were excluded if validation of step counting was not an aim, the index device was not a consumer wearable or smartphone, an accelerometer was used as the criterion measure (laboratory studies), a non-research-based accelerometer was used as the criterion measure (semifree-living or free-living studies), the full text was not made available or it was published in the grey literature (eg, congresses, conferences or meeting abstracts). The systematic literature review process was conducted by WJ, PBJ and PMG. Title/abstract screening and full-text screening were completed by two independent reviewers and confirmed by a third independent reviewer using Covidence software (Veritas Health Innovation). 26 Data were extracted using a standardised template developed by the consortium, allowing for the aggregation of study data under the following headings: sample size included/analysed, sex distribution, type of population, body mass index (BMI), height, weight, age, condition type (laboratory; semifree-living; free-living), criterion measure (description, configuration and placement), index measure (description, configuration and placement), testing protocol, signal processing, data synchronisation, statistical analysis, results and conclusions. The studies were divided into three main condition categories based on the following definitions:

Laboratory: Highly controlled methodologies that involved participants completing gait tasks at controlled or self-selected speeds. This criterion includes studies that require participants to complete gait tasks in indoor and/or outdoor environments, and may require individuals to ambulate on any surface or involve ascent or descent of stairs.

Semifree-living: Semicontrolled methodologies that involve participants completing 'simulated' activities of daily living for the purpose of replicating 'free-living' conditions (eg, sweeping, cooking, computer use, etc).

Free-living: Uncontrolled methodologies that involve participants wearing the index device during 'normal' daily life, outside of a controlled laboratory or simulated environment.

To facilitate evaluation of the current validation protocols, the risk of bias for each article was evaluated under the following headings using the Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool: patient selection, index measure, criterion measure, and study flow and timing. 27 The methods for the QUADAS-2 evaluation are presented in online supplemental appendix 2.

Additional specific searches of the wider literature were then conducted to help identify and find solutions to address factors that introduce bias during the validation of these devices. For example, identify the optimal walkway length to achieve steady-state walking. The following section presents

the current state of knowledge based on the evidence synthesised during literature review process and the INTERLIVE members evidence informed expert opinion.

The search strategy for this systematic literature review identified 2897 potential articles for inclusion. After the removal of 704 duplicates, 2193 articles were eligible for title and abstract screening. Following this, 174 articles met the criteria for full text screening. Eighty-nine studies were excluded, resulting in a total of 85 articles eligible for the final dataset. Figure 1 illustrates the various stages of the article screening process. Twenty-nine studies investigated the validity of a smartphone-based step counting application, 65 studies investigated a consumer wearable activity tracker, while three studies investigated the validity of an item of smart clothing. Online supplemental appendix 3 (laboratory), online supplemental appendix 4 (semifree-living) and online supplemental appendix 5 (free-living) present the validation methodologies used across the different studies. The risk of bias assessment for each study is presented in online supplementary appendices 6–8. Within our systematic literature review, 98 unique consumer wearables and smartphones were identified, distributed across 50 different brands. From Apple, 11 models of smartphone and 1 smart-watch were included. Nine unique devices were included from Samsung, eight from Fitbit, five from Jawbone, five from Sony, four from Garmin, three from Withings, three from Huawei and two from Omron. An additional 48 unique devices were identified across 45 different brands. The most studied brand was the Fitbit, followed by Apple, Jawbone and Garmin.

PRISMA flow chart of the systematic review process. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

From the synthesised data obtained during the literature review process, the INTERLIVE members identified a set of six key validation protocol and reporting domains: (1) target population, (2) criterion measure, (3) index measure, (4) validation conditions (laboratory, semifree-living and free-living), (5) processing and (6) statistical analysis. A set of key variables of interest were then identified based on an evidence informed expert opinion (figure 2). The following section synthesises the evidence obtained during the statement process and outlines the current state of knowledge regarding validation of consumer wearable and smartphone step counters.

Six domains and corresponding set of key variables of interest which were identified as being of importance when considering validation of consumer wearable and smartphone step counters.

It is well understood that gait characteristics are significantly influenced by factors such as physical function, 28 BMI, 29 30 age, 31 and sex. 32 33 For example, those with Parkinson's disease walk with a short-stepped, narrow based, shuffling gait, when compared with healthy adults. 34 Additionally, children and adolescents with obesity have different kinematic and kinetic gait patterns when compared with those with normal weight. 29 Importantly, if step counting algorithms are developed using solely 'typical' gait, the accuracy of wearable and smartphone step counters may be influenced by these different 'atypical' gait characteristics. The impact of this has been demonstrated by studies that have highlighted the increased step counting error of these devices at slower walking speeds and in clinical cohorts. 35–38 To ensure that the validation of the device is representative of the general population, best practice validation should include those with heterogeneous physical function, age and sex. To facilitate this, we propose that those interested in the validation of these step counting devices for use within the wider population should consider separate evaluation of the device within those with 'typical' gait characteristics ('basic' protocol) and 'atypical' gait characteristics ('advanced' protocol). Figure 3 provides an overview of the multi-stage nature of the consumer wearable and smartphone step counter validation protocols.

The multistage protocols for the best-practice validation of consumer wearable and smartphone step counters. We recommend a minimum validation in those with typical gait (basic protocol), across laboratory (step 1) and semifree-living (step 2).

For the basic protocol, recruitment should span three groups: children (<12 years), adolescents/adults (13–64 years) and older adults (≥65 years). The focus of the basic protocol is to validate the step counting device within a 'typical' gait population, capturing data from individuals with a range of gait characteristics (eg, speed, cadence, etc). As evidence suggests that gait characteristics stabilise with puberty (12–14 years) 39 and are altered in older adults, 40 we recommend to separate recruitment of children, adolescents/adults and older adults. This will help to identify how the device performs across the spectrum of age, among those without 'atypical' characteristics that have been shown to affect step counting.

For the advanced protocol, recruitment should focus on validation in populations with 'atypical' gait. The focus of this protocol is to introduce factors that are known to negatively affect gait kinematics or wearable/smartphone-based step counting. This will help ensure that the validation process considers the heterogeneous nature of the general population. Examples of the types of populations that should be considered (but not limited to) are those with morbid obesity, 29 falls-risk, 41 lower-limb amputations 42 and neurological conditions. 43 The validation protocol should be completed within the specific population of interest, and the demographic characteristics of the group should be reported.

Across all studies obtained during the systematic literature review process, the average (range) sample size was 45 (5–592) participants, while authors did not typically complete a formal sample size calculation. Of the 11 studies that conducted formal sample size calculations, the average (range) sample size necessary to reach a power of 0.80 was 42 (16–118) participants. If the focus of the study is to conduct hypothesis testing about a predefined minimal level of accuracy, we advise that a sample size calculation based on previously published or pilot testing mean and SE of the differences between the two devices and predefined clinical maximum allowed difference should be conducted, using the methodologies outline by Lu et al . 44 If sufficient data are not available or this is not the primary focus, based on the average sample size observed in the synthesis of studies which conducted sample size analyses included within the systematic literature review, the observed mean differences and SD of the differences, we advise a minimum of 15 participants for each specific validation group, totalling ≥45 participants, accounting for potential drop-out or data loss.

An important consideration during the validation of any device is the validity of the criterion measure to which the index device is being compared. If the criterion measure is not sufficiently valid, then criterion standard bias may be present, 45 limiting the value that may be gained from such a study. In the context of step counting, across the previously published literature, a total of seven different criterion methods were used. The most common method was a research grade wearable device (n=35 (41%)). This was followed by video recording with one observer (n=23 (27%)), and visual observation with one observer (n=20 (24%)). Only eight (9%) studies used the most rigorous method of video analysis with at least two observers. The most common method under laboratory and semifree-living conditions was video recording, while free-living studies typically used a research grade wearable device (eg, Actigraph, Stepwatch, etc).

Based on the current available evidence, we advise the use of video with two observers for a gold-standard criterion measure of step count. While the use of video recording within a laboratory or semifree-living setting is feasible, the feasibility of using video within a free-living scenario needs to be considered. Previous studies have investigated the feasibility of wearing a body worn camera for 10–14 hours during free-living. 46 These authors concluded that compliance with the protocol was 87.5%, with participants reporting that they had not changed their behaviour as a result of wearing the device and were not concerned by the amount of wear-time. 46 From the evaluator's perspective, feasibility is compromised by the processing time that video recording requires. For instance, a sample size of 45 participants, recorded during 10–14 hours implies 450–630 hours of video to be analysed by two independent observers, which means 4.8–15.8 months of full-time work (8hours/day × 5 days/week × 4 weeks/month) multiplied by two persons, exclusively devoted to counting steps from video recorded. Noteworthy is that these calculations are done for the scenario of a single day (ie, 10–14 waking hours) of video record only, when ideally, free-living

validations should be done for a complete week (as usually done in accelerometer-determined PA studies 47 to cover the interday variability of human activity behaviour. Importantly, the systematic literature review highlighted only one study that has leveraged video recording in free-living conditions. 21 This investigation 21 leveraged a waist mounted camera to obtain a gold-standard measurement of step count during free-living activity, over a 24-hour period.

While this approach is undeniably challenging to incorporate into the validation design, it must be acknowledged that currently, no other method has shown sufficient validity to be considered as an alternative gold standard. Despite being commonly used in free-living validation of consumer devices, studies have highlighted that research grade wearable devices are highly sensitive to atypical gait 35–38 and sensor mounting location. 35 For example, it has been highlighted that the wrist mounted Actigraph wGT3X+ has a high MAPE during slow walking at 1.6 km/hour (wrist 47%) and 3.2 km/hour (wrist 22%). Furthermore, the hip mounting location increased the MAPE to 82% (1.6 km/hour) and 24% (3.2 km/hour), respectively. 35 Similarly, the ActivePAL device has been shown to demonstrate increasing step counting error during slower walking speeds. 48 49 However, this is a rapidly evolving field and despite these current limitations, such approaches are beginning to demonstrate potential. In the aforementioned study, 21 the StepWatch demonstrated an MAPE of 4% when compared with free-living video in 12 healthy adults with typical gait, highlighting the potential of this approach. However, prior to incorporating such approaches into the validation, we advise that further validation of these potentially 'equivalent' criterion methods should be explored, using the validation protocol presented within this statement. The alternative approach should only be considered 'equivalent' to the video gold standard if it achieves an $MAPE \leq 5\%$ and there is no evidence of proportional or fixed bias across laboratory, semifree-living, and free-living conditions. In addition, these promising findings should be replicated in future cross-validation studies and if similarly low errors are observed, StepWatch or any other alternative device (accelerometers, insoles, computer vision) could be used as an 'equivalent' criterion. Any margin of error should then be reported alongside the study methodology. It is not deemed acceptable to only assess the validity of an alternative criterion device under laboratory or semifree-living, without considering free-living conditions.

An important aspect to consider is the wear location of the wearable or smartphone device during the validation. As the focus of these consumer devices is for unobtrusive use by the general population, it is important for manufacturers and users to ensure that the device is validated in an ecological wear location—the location in which consumers typically use the device. For example, individuals frequently carry smartphones in their pocket, 50–52 bag 51 53 or hands, 51 54 with a cross sectional study highlighting that 60% of females carry their phone in their bag and 60% of men carry their phone in their pocket. 55 Furthermore, activity trackers are normally worn on the wrist. 56–58 Conversely, smartphones and wearable activity trackers are not usually attached to the waist, 16 59–61 hip 62 63 and chest, 17 as was observed in several validation studies. Importantly, it has been shown that step counting error is impacted by wear location, with hip-based devices demonstrating a higher accuracy than wrist-worn devices, 64 while smartphone accuracy changes based on the device, location (waist, arm, hand or bag) and nature of the activity (running or walking). 59 65 Therefore, it is important to ensure that there is consistency in wear locations within an individual study. We advise that the wear location is specified for each laboratory and semifree-living evaluation. Failure to do so may result in a high degree of inter-subject variability, impacting statistical power and limiting the conclusions that can be drawn from the validation process.

The primary function of consumer wearable and smartphone step counters is to provide objective data pertaining to PA, outside of the laboratory environment. However, the systematic literature review highlighted that no single study extensively investigated the validity of the consumer device across controlled laboratory and uncontrolled free-living conditions (online supplemental appendices 3–5). Of the 85 studies included, 57 examined validity using constrained laboratory validation conditions, nine used semifree-living conditions, while 30 used free-living conditions. Two studies used both laboratory and semifree-living conditions, nine used laboratory and free-living conditions, while only one study used semifree-living and free-living conditions. Therefore, the

validation protocol proposed within this expert statement has been designed to focus on aspects that have been shown to influence wearable and smartphone step counting algorithms, through a three-stage process.

Laboratory evaluation (stage 1) aims to evaluate the ability of the device to accurately measure steps during self-selected walking and running speeds. This stage focuses on aspects that have previously been shown to negatively affect step counting accuracy, such as short bouts of walking/running, discontinuous walking/running, change of direction, stair climbing and cycling. 66–68 As previous research has highlighted that 1–7 m is required to enter steady-state-gait, 69 it is important to ensure that the walkway is sufficiently long. Therefore, as the typical gymnasium is approximately 30 m long (a typical basketball court is 28m × 15m), we advise that the laboratory walking/running trials are conducted using a 30 m long track. Besides validity, an additional evaluation of the within-device precision (reliability) should be investigated during a steady state treadmill walking test. This evaluation should involve two repeated 1 min steady state walking tests, completed at consistent self-selected walking speeds. Previous kinematic research has indicated that footwear 70 and ground surface type 71 72 can significantly influence gait kinematics and kinetics. However, a study investigating the impact of these changes on step count accuracy in consumer wearables demonstrated that floor surface (grass, gravel, linoleum, tarmacadam, tile and ramp) did not significantly impact step count accuracy. 73 Importantly, the authors demonstrated that the step count error may be increased during stair walking (ascending and descending). 73 Therefore, we advise that the laboratory validation is conducted on both flat surfaces and during stair-walking. If the investigators are particularly interested in the validation of the device on alternative surfaces we advise that the laboratory protocol is conducted on relevant alternative surfaces (grass, gravel, linoleum, tarmacadam, tile, incline/decline, etc) to ensure ecological validity.

The semifree-living evaluation (stage 2) aims to examine the performance of the step counters during simulated sedentary, household and exercise-based activities of daily living. It is well established that activities of daily living frequently result in the false detection of steps by wearable and smartphone devices. A previous study 66 demonstrated that non-stepping activities such as deskwork, taking an elevator, taking a bus journey, driving an automobile, washing and drying dishes, functional reaching tasks, indoor cycling, outdoor cycling and indoor rowing often result in the detection of false steps. While there is evidence 74 suggesting that monitors may underestimate steps by 35%–64% during household tasks. The proposed sedentary, household and exercise-based tasks are among the most common PAs in both youth and adulthood, covering a wide spectrum of the human movement pattern. 75 76 For instance, we include different positions (lying, sitting and standing), fine and gross motor control tasks such as writing or rowing exercises respectively and the combination of steady state situations (eg, standing and talking) with locomotion activities (eg, simulated shopping). We also include a considerable range of intensities according to the compendium of PAs, which is relevant in terms of step detection since some wearables have demonstrated a weak performance in low intensity activities. 35 Activities that involve arms and hands movements have registered a great number of false steps, and therefore, we propose the inclusion of arm-movement activities not including steps (eg, washing dishes) and including steps (eg, vacuuming). 57 66 77 Ultimately, the number of potentially valuable simulated activities to include is endless; however, from a pragmatical point of view, the protocol we propose includes a selection of relevant activities that can be performed in less than 45 min. The incorporation of a free-living validation component will serve to ensure that a wide range of activities that have not been considered in laboratory conditions are included.

The free-living evaluation (stage 3) aims to capture a wide array of activities of daily living ‘in the wild’, accounting for the differences observed between controlled and uncontrolled activities. Evidence suggests that individuals typically have different gait characteristics during laboratory and free-living assessments. 28 The impact of this is highlighted by the increased wearable and smartphone step counting error observed during free-living conditions, 21 when compared with laboratory conditions. 20 We advise that during this stage, participants are encouraged to engage in their usual activities of daily living, to quantify measurement error in an unconstrained

environment, over a period of at least 24 hours, but ideally 1 week. A week long recording period, resulting in a minimum of 3–4 days of recorded time is optimal to capture interday and weekday-weekend day variation.

When comparing a method of measurement to a gold-standard criterion, an important aspect that should not be overlooked is the development of a precise synchronisation methodology to align the data captured from the index and criterion. While the methods used to synchronise the measures may seem trivial to some, if not done correctly, any discrepancy may introduce error, biasing the results. For example, the step count on a device may be 'contaminated' by activity that took place before or after the activity recorded by the criterion measure.

We advise that adequate synchronisation is achieved by requiring participants to stand still at the beginning and the end of all trials while the step count and video time are noted. This pragmatic approach should be used across laboratory, semifree-living, and free-living conditions, and may be specific to the index device of interest. For example, some devices may allow step count to be recorded from the device interface, while others may require synchronisation with a smartphone before it can be viewed. 66

If an alternative 'equivalent' criterion measure to video is being used, then the time-stamp and/or step count should be noted at the beginning and the end of each trial for both measures. In addition, the synchronisation method should be reported in detail within the validation report to ensure transparency and replicability.

If video is used as the criterion measure, to ensure gold-standard validity, ≥ 2 observers should view each video in real-time and count the number of steps using a manual counting device. During periods of uncertainty, where it is not clear if a step took place, the video should be reviewed in slow motion. The systematic literature review highlighted eight studies that used ≥ 2 observers counting the number of steps. 37 78–84 The observers should be sufficiently trained to ensure excellent agreement between reviewers and the interobserver reliability should be reported. 37 80 82 If an equivalent criterion method is being used, the data processing methods that have been described within the previously published validation study should be followed.

An important aspect to consider in the validation of these step counting devices is what constitutes a 'step'. Typically, a step is defined in biomechanical terms as the period during the gait cycle between two consecutive heel strikes 85 or in more general terms as 'the act of putting one leg in front of the other in walking or running'. 86 However, these definitions are not suitable for this context as they are either too constrained (biomechanical definition) or do not provide sufficient guidance for deciding what constitutes a step within this context (locomotion definition). Furthermore, these definitions disregards activities of daily living that involve steps but without a locomotive purpose, such as vacuuming or hanging out clothes. Therefore, for the purpose of validating wearable and smartphone step counters that are designed for the purpose of quantifying PA and energy expenditure, the INTERLIVE network has defined a step for the as follows:

'The act of raising one foot and putting it down in another spot, resulting in the displacement of the centre of mass'.

This definition will allow for step counting during continuous locomotion (walking/running), as well as during intermittent gait, common during activities of daily living.

In order to thoroughly consider the validity of an index device when compared with a criterion measure, agreement between the two devices should be evaluated. 87 Agreement provides details about how accurate a measurement tool is when compared with a reference standard. 88 A systematic literature review of statistical methods employed in medical instrument validity studies highlighted that a large proportion of studies have used inappropriate statistical methods. 89 In line with these conclusions, many of the studies included within our systematic literature review

inappropriately used measures of relative measurement reliability (ie, intraclass correlation coefficient, kappa, etc) (n=37 (44%)), correlation (ie, Pearson, Spearman, etc) (n=34 (40%)), and comparison of means (n=33 (39%)).

To thoroughly investigate validity, we advise the use of Bland-Altman limits of agreement analysis and MAPE. Bland-Altman analysis is widely accepted as the most appropriate tool in assessing agreement within medical validation studies, providing a measure of the agreement between the two measurements.^{87 89} This method was the most common approach seen across the previously published studies (n=43 (51%)). To investigate the presence of any proportional or fixed bias, least-products regression should be used as part of the Bland-Altman analysis.⁹⁰ To facilitate comparison between devices and testing conditions, an absolute measure of error, such as MAPE should be used. MAPE is the average of the absolute percentage of error of a tool and is commonly used to describe the error of a prediction model,⁹¹ and was used by 28 studies within our review (33%). Together, these statistical approaches provide a means to thoroughly evaluate the validity of the wearable and smartphone step counters and help identify the presence of any measurement bias.

The within-device precision (reliability) of the index device should be investigated by comparing the total number of steps obtained during the repeated 1 min steady state treadmill walking tests, using intra-class correlation coefficient with 95% CIs.

When considering what constitutes a valid device, readers should consider the results within the context of the potential use case. Specifically, if an activity tracker is to be used as an outcome measure within a clinical trial or as an alternative gold-standard measurement tool for step counting, we recommend that the device should demonstrate an extremely low level of measurement error (MAPE $\leq 5\%$) and bias, within the clinical population of interest. Conversely, if the device is being validated for use by the general population, then we recommend that a marginally higher error (MAPE $\leq 10\text{--}15\%$) may be acceptable. Therefore, if the goal of the evaluation is not to conduct hypotheses testing and a formal sample size calculation has not been performed, binary conclusions about the validity should not be made. Rather, the level of measurement error and bias should be openly reported, taking into consideration all the proposed validation conditions (laboratory; semifree-living; free-living), to facilitate contextual interpretation.

Based on the available evidence outlined within this statement, INTERLIVE recommends that wearable and smartphone step counter manufacturers seek to validate their devices using standardised and transparent methodologies. We advise that manufacturers engage in a multistage validation protocol. Broadly, this protocol examines the validity of these step counter devices under laboratory (stage 1), semifree-living (stage 2) and free-living conditions (stage 3), across populations with 'typical' (basic protocol) and 'atypical' (advanced protocol) gait characteristics. The basic protocol should be used by those interested in demonstrating the validity of step counting devices within populations that do not demonstrate 'atypical' gait characteristics known to influence wearable and smartphone based step counting. If the goal of the evaluator is to demonstrate the validity of the device across the wider population in those with known 'atypical' gait characteristics, the advanced protocol should also be used. Figure 3 provides an overview of the multistage nature of the consumer wearable and smartphone step counter validation protocols. The proposed framework sets out to ensure that the recruited target population is fit-for-purpose, addressing many of the sources of bias identified during the review of the literature. Table 1 presents a detailed best-practice validation protocol, while table 2 presents a check-list designed to facilitate validation protocol planning. In addition, a comprehensive protocol reporting sheet for use by both research institutions and industry is presented within table 3 to facilitate standardised and transparent data sharing. INTERLIVE advises that wearable and smartphone step counter manufacturers seek to achieve the basic and advanced multistage validation protocols, completing the laboratory, semifree-living and free-living validation (stages) across populations with 'typical' (basic protocol) and 'atypical' (advanced protocol) gait characteristics. However, in the scenario where it is not possible to perform the whole validation protocol, at a minimum, the basic laboratory (stage 1) and semifree-living (stage 2) validation, within a 'typical' gait population should be achieved.

The proposed best-practice protocols for the validation of wearable and smartphone step counting devices

3 min zig-zag walk test (30 m walkway - three cones)

3 x stair test ascent/descent (minimum 12 step staircase)

2×1 min steady state treadmill test (consistent, self-selected walking speed).

3 min zigzag run test (30 m walkway - three cones)

simulated sleeping (Get into bed, after 1 min roll over, after 2 min check time on smartphone/wearable, after 3 min get up)

computer use (both typing and mouse use)

simulated video-game (game console controller)

folding laundry (no steps while standing)

simulated washing/drying dishes (washing for 1.5 min and drying for 1.5 min)

simulated shopping (pick up two shopping bags, walk 10 m, empty bag and put items in cupboard—perform task once)

Least products regression of the differences against the means

Intraclass correlation with 95% CIs (calculated for the 2×1 min treadmill test only)

If the evaluator decides not to perform the whole validation protocol, there is a minimal protocol to be followed. At a minimum, the validity should be investigated in those with typical gait (basic protocol), across laboratory (step 1) and semifree-living (step 2). Observers should determine which participants are deemed suitable for the different components of the assessment protocols.

*Any device or method which has been demonstrated to possess less than 5% measurement error using the laboratory, semifree-living, and free-living validation protocols detailed below. Additionally, this should be specific to the population of interest. For the case of free-living validation, the INTERLIVE network is aware that video recording and 'manual' step counting by two independent evaluators over ≥ 24 hours recording period (a 1-week recording period (minimum 3–4 complete days) is optimal to capture interday and weekday-weekend variation) is extremely costly in time and resources and therefore likely not feasible for many. Therefore, we feel that this field needs to move forward, developing and validating alternative methods such as insole sensors that after validation and cross-validation could be used as a new and more feasible gold-standard method for free-living validation protocols.

BMI, body mass index; INTERLIVE ® , Intelligent Health and Well-Being Network of Physical Activity Assessment; LoA, limits of agreement; MAPE, mean absolute percentage error.

Check-list of items to be considered during the validation protocol development for consumer wearable and smartphone step counters

*Any device or method which has been demonstrated to possess less than 5% measurement error using the laboratory, semifree-living, and free-living validation protocols detailed below.

Additionally, this should be specific to the population of interest. For the case of free-living validation, the INTERLIVE network is aware that video recording and 'manual' step counting by two independent evaluators over ≥ 24 hours recording period (a 1-week recording period (minimum 3–4 complete days) is optimal to capture interday and weekday-weekend variation) is extremely costly in time and resources and therefore likely not feasible for many. Therefore, we feel that this field needs to move forward, developing and validating alternative methods such as insole sensors that after validation and cross-validation could be used as a new and more feasible gold-standard method for free-living validation protocols.

†If an 'equivalent' criterion is used, the device should be used as per the methods described within the validation study.

INTERLIVE[®], Intelligent Health and Well-Being Network of Physical Activity Assessment; LoA, limits of agreement.

Minimum required reporting sheet for standardised and transparent data sharing

*Any device or method which has been demonstrated to possess less than 5% measurement error using the laboratory, semi free-living, and free-living validation protocols detailed below. Additionally, this should be specific to the population of interest. For the case of free-living validation, the INTERLIVE network is aware that video recording and 'manual' step counting by two independent evaluators over ≥ 24 hours recording period (a 1-week recording period (minimum 3–4 complete days) is optimal to capture interday and weekday-weekend variation) is extremely costly in time and resources and therefore likely not feasible for many. Therefore, we feel that this field needs to move forward, developing and validating alternative methods such as insole sensors that after validation and cross-validation could be used as a new and more feasible gold-standard method for free-living validation protocols.

BMI, body mass index; INTERLIVE[®], Intelligent Health and Well-Being Network of Physical Activity Assessment; LoA, limits of agreement; MAPE, mean absolute percentage error.;

This INTERLIVE expert statement strives to provide actionable and unambiguous recommendations for those interested in the thorough and transparent validation for consumer wearable and smartphone step counters. Based on the best available evidence, a detailed validation protocol has been presented under the domains of target population, criterion measure, index measure, validation conditions (laboratory, semifree-living and free-living), processing and statistical analysis.

It is clear that past approaches used by industry to validate wearable and smartphone step counters are not fit-for-purpose. While early attempts by the CTA to develop validation standards were a step in the right direction, the CTA validation approach did not address many of the potential sources of bias that have been highlighted above and has not been widely adopted. 19 As such, the decision on device validation lies with the manufacturer, with no regulatory or certification standards currently in place to ensure transparency and scientific rigour. This means that there is a high degree of heterogeneity across the methodologies used to validate these devices, with no means for interested stakeholders to objectively compare devices. A rigorous, standardised and transparent validation process should be the mutual interest of manufactures, scientific institutions and consumers in order to judge whether these devices are useful and perform with satisfactorily low measurement error. The approach presented within this statement will help progress towards the goal of 'evidence-based marketing claims', ensuring that wearable technology can be used safely and to its full potential. 18 It is intended that the recommendations outlined within this statement provide an evidence-based means for manufacturers to ensure that their devices are validated to an acceptable standard, using rigorous and transparent scientific methods. As a next step, INTERLIVE plans to provide best-practice recommendations for validation of wearable and smartphone based energy expenditure and maximal oxygen consumption (VO₂ max) estimations,

contributing to the development of accepted standards for the validation of wearable and smartphone PA tracking devices.

If the standardised validation of these devices can be realised, there are many potential positive implications for all stakeholders. Manufacturers will be able to transparently communicate the value of their devices to consumers. Furthermore, transparency will drive innovation, as the industry strives to achieve improved validity. General consumers will be able to compare the validity of potential devices, ensuring that they can make an informed decision about the most appropriate device for them. Similarly, this will improve transparency for healthcare providers who are interested in incorporating these devices into every day clinical practice, facilitating evidence based digital healthcare. Finally, scientific researchers will be able to understand the validity of these devices, facilitating progress in the realms of 'passive' digital phenotyping and 'active' PA behavioural change research. As such, we urge all stakeholders to engage in the development and validation of these devices to adhere to these best-practice recommendations.

The thorough validation of these devices does not come without challenges. The systematic literature review process highlighted a wide range of criterion methods used to count steps. Despite this, we advise that there is currently not sufficient evidence to support the use of alternatives to video recording as the gold-standard measurement of step count. However, as mentioned above, there are significant challenges associated with the use of video analysis within the free-living environment. Therefore, we recommend that researchers and manufacturers should focus on the development and thorough validation of alternatives that may improve the practicality of free-living validation. Potential approaches that are beginning to demonstrate promise are the use of research grade accelerometers placed at the ankle or foot (for closer relation to the foot stepping vs the usual placements of hip or wrist), 21 92 insoles 93 and machine learning-based video analysis. 60 If proven to be sufficiently accurate when compared with video-based step count, these approaches may serve to improve the feasibility of free-living validation studies.

This INTERLIVE expert statement provides an evidence-informed best-practice consumer wearable and smartphone step counter validation protocol. The systematic literature review included within, highlighted a high degree of heterogeneity between previously published methods, with many studies failing to address key sources of validation bias. The INTERLIVE consortium recommend that the proposed validation protocol should be used when considering the validation of any consumer wearable or smartphone step counter to overcome the main identified issues. The consortium also provide guidance for future research activities, highlighting the imminent need for the development of feasible alternative 'gold-standard' criterion measures for free-living validation. Adherence to this validation standard will help ensure methodological and reporting consistency, facilitating comparisons between consumer devices and the amalgamation of standardised open datasets. This will ensure that manufacturers, consumers, healthcare providers and researchers can use this technology safely and to its full potential.

Twitter: @Will_Johns10, @moritz_schumann, @Ulf_Ekelund

Contributors: All authors were involved in the development and drafting of the expert statement. All authors have read and approved the content of the manuscript.

Funding: PBJ is supported by the Portuguese Foundation for Science and Technology (SFRH/BPD/115977/2016). WJ is partly funded by Science Foundation Ireland (12/RC/2289_P2). PMG and FBO are supported by grants from the MINECO/FEDER (DEP2016-79512-R) and from the University of Granada, Plan Propio de Investigación 2016, Excellence actions: Units of Excellence; Scientific Excellence Unit on Exercise and Health (UCEES); Junta de Andalucía, Consejería de Conocimiento, Investigación y Universidades and European Regional Development Funds (ref. SOMM17/6107/UGR). JMM is partly funded by Private Stiftung Ewald Marquardt für Wissenschaft und Technik, Kunst und Kultur. UE and JS are partly funded by the Research Council of Norway (249932/F20). AG is supported a European Research Council Grant (grant number 716657). ELS is supported by TrygFonden (grant number 310081). This research was partly funded

by Huawei Technologies, Finland.

Supplemental material: This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.