SURVEY

# Associating Physical Function and Capacity Tests to Free-Living Sensor Data: A Systematic Review on Technology and Methods.



**Accepted:** 25 January 2026
**Revised:** 12 December 2025
**Received:** 11 July 2025

# Associating Physical Function and Capacity Tests to Free-Living Sensor Data: A Systematic Review on Technology and Methods.

SARA CARAMASCHI, Sustainable Digitalisation Research Centre, Department of Computer Science and Media Technology, Malmö University, Sweden
DARIO GHEZZI, National Research Council, Information Science and Technologies Institute Alessandro Faedo, Italy
CARL MAGNUS OLSSON, Sustainable Digitalisation Research Centre, Department of Computer Science and Media Technology, Malmö University, Sweden
FILIPPO PALUMBO, National Research Council, Information Science and Technologies Institute Alessandro Faedo, Italy
DARIO SALVI, Sustainable Digitalisation Research Centre, Department of Computer Science and Media Technology, Malmö University, Sweden

Physical function and capacity tests are widely used for assessing health across various clinical conditions. However, traditional assessments may not accurately capture real-world health conditions reliably and frequently. Sensors, smartphones and wearable devices offer the potential to bridge this gap by collecting data in everyday life that may better reflect participants' physical capabilities, and could be used to predict clinical outcomes and the performance of physical tests. However, there is a lack of comprehensive reviews and consensus in the field. This work reviews the literature on passively collected data from digital health technology in relation to physical function and capacity tests and informs future investigations in this domain. A systematic literature search was conducted following the PRISMA guidelines on 3 databases. Our analysis identifies cardiovascular and neurodegenerative diseases as the most frequently studied conditions, and wearables embedding inertial sensors as the most common device type. Most studies rely on one week-long data collection. Associations between physical test outcomes and metrics such as step count and activity intensity show correlations as high as 0.89 when machine learning is introduced. This review provides a comprehensive summary of current research on the use of digital health technology in free-living conditions and the clinical significance of data when associated with physical tests.

CCS Concepts: • **Human-centered computing** → *Mobile devices*; *Mobile phones*; • **General and reference** → **Surveys and overviews**; • **Hardware** → *Sensors and actuators*; **Sensor applications and deployments**.

Additional Key Words and Phrases: Digital health, Passive monitoring, Physical activity, Physical tests, Free-living conditions, Smartphones, Telehealth, Wearables

Authors' Contact Information: Sara Caramaschi, Sustainable Digitalisation Research Centre, Department of Computer Science and Media Technology, Malmö University, Malmö, Sweden, sara.caramaschi@mau.se; Dario Ghezzi, National Research Council, Information Science and Technologies Institute Alessandro Faedo, Pisa, Italy, dario.ghezzi@isti.cnr.it; Carl Magnus Olsson, Sustainable Digitalisation Research Centre, Department of Computer Science and Media Technology, Malmö University, Malmö, Sweden, carl.magnus.olsson@mau.se; Filippo Palumbo, National Research Council, Information Science and Technologies Institute Alessandro Faedo, Pisa, Italy, filippo.palumbo@isti.cnr.it; Dario Salvi, Sustainable Digitalisation Research Centre, Department of Computer Science and Media Technology, Malmö University, Malmö, Sweden, dario.salvi@mau.se.


ACM 2637-8051/2026/2-ART
https://doi.org/10.1145/3797893

## 1 Introduction

Advances in wearable sensors and mobile devices have opened up new possibilities for the continuous monitoring of human physical activity, heart rate, and sleep patterns, among other health-related metrics [1]. The analysis of mobility and physiological data collected passively during everyday life (real-world or free-living conditions) by commercial wearable devices, smartphones, and related sensors (digital health technology (DHT)) has the potential to introduce new clinical evaluation methods that could improve healthcare efficiency and complement traditional clinical assessments [2]. However, to use these data for clinical purposes, they must be associated with clinically meaningful indicators such as the results of validated and objective physical tests. Standard physical tests are common practice when assessing and monitoring patients' physical function and capacity. Among the most used ones are the 6-Minute Walk Test (6MWT), which measures the distance that the patient can walk in 6 minutes at a natural pace, and the Timed Up and Go Test (TUG), which investigates the time required to stand up from a chair, walk three meters, turn around, go back and sit down on the chair again. These tests evaluate whether a person can perform certain basic or instrumental activities of daily living [3, 4] or the extent to which the person can sustain physical effort [5].

Despite their widespread use in clinical areas such as cardiopulmonary disease, neurological disorders, and geriatrics, these simple clinical tests can still be burdensome as they require the presence of clinicians and require patients to travel to the clinic. The associated costs and inconvenience for patients and healthcare providers lead to infrequent testing and may therefore fail to capture symptom variability. In addition, the clinical tests may not always reflect patients' actual physical function in everyday settings. Several studies highlight such limitations, noting that they lack comprehensive information about how an individual's mobility translates to real-world conditions [6]. Bansal et al. [7] furthermore argue for shifting the focus from purely capacity-based measures — especially in populations with positive test outcomes (e.g., gait speed $\geq 0.49\,m/s$ or 6-Minute Walk Distance (6MWD) $\geq 204\,m$) — to observing community ambulation via wearable devices to capture post-stroke ambulation better. Similarly, Zajac et al. [8] emphasize the importance of continuous data collection in monitoring individuals with Parkinson Disease (PD) rather than relying solely on test-derived measurements.

Real-world physical activity is becoming increasingly relevant in clinical trials as a complement to the 6MWT [9], for example, in pulmonary arterial hypertension, where physical activity is associated with quality of life and clinical outcomes. This is an example of how obtaining equivalent or complementary information to physical tests may support clinicians in decision-making and highlights the importance, meaning, and use of data collected in real-world conditions.

In a review of wearable devices used in cancer [10], 25 studies were found where accelerometers and their variants were used. This was primarily done to measure physical activity, circadian rhythm, sleep, and skin temperature. Of those studies, 9 reported correlations between wearable data and patient-reported outcomes such as quality of life, symptoms and mental health. While the review indicates that data from wearables can be related to clinical outcomes, at least in cancer, the outcomes that were used were self-reported and therefore somewhat subjective, thus leaving a need to explore the link with more objective health indicators.

A systematic review on the influence of wearable devices in chronic diseases identified 30 studies where these devices were used as a means to improve healthcare outcomes [11]. The most studied chronic conditions were diabetes, Parkinson's disease and chronic lower back pain. The use of wearables was often associated with outcomes such as pain, quality of life, and physical function. The review focused on the positive impact of the use of the devices, mostly as a result of behavioural changes caused by the information provided by the devices and their companion apps. While providing useful information, the review does not discuss the clinical significance of the data produced by those devices.

Other reviews on sensors and wearable devices [12, 13, 14] report mostly technical and usability aspects. These include form factor, sensing capabilities, as well as hardware and software characteristics rather than the clinical

relevance of the data collected by such devices. Reviews also exist that investigate the use of passively collected data, such as the one by Giurgiu et al. [15] for adults and by Bernaldo et al. [16] for stroke patients. However, none of these reviews assess the association between physical function and capacity tests and passively collected data.

Subsequently, while there is growing interest in the use of DHT in daily life for healthcare purposes [17, 18], comprehensive reviews of the clinical validity of continuous passive monitoring are still scarce. To the best of our knowledge, no review has examined the relationship between data collected in free-living conditions and physical tests such as the 6MWT and the TUG.

### 1.1 Objectives and research questions

This review aims to provide a comprehensive summary of current research on the use of sensors in free-living conditions and their clinical significance. Based on this, we additionally identify guidelines and methodological considerations relevant for future studies that want to investigate people's physical performance, function and capacity through the use of widely available technology.

The following are the research questions addressed in this systematic review:

- RQ1 Is it possible to associate data collected by DHT in a free-living environment with standard physical tests?
- RQ2 Which are the most common DHT used to collect passive data from users?
- RQ3 Which metrics are extracted from the data during daily living, and how are these metrics associated with physical function and capacity tests?
- RQ4 What are the most common characteristics of the experimental protocols in the reviewed studies?

The remainder of this review is structured as follows: Section 2 provides a detailed background on the review topic. Section 3 describes the search strategy and analysis methods. Next, Section 4 presents the analysis and findings from the reviewed articles according to each theme addressed. Finally, Section 5 discusses the findings in relation to the proposed research questions, and Section 6 concludes the review by outlining the key insights and suggestions for future research.

## 2 Background

In scientific literature, numerous tests are described to evaluate physical health through the measurement of function and capacity [19]. By physical function, we mean that set of aspects linked to the neuromotor sphere through which it is possible to perform "basic mobility skills" as daily activities and tasks, such as standing up, sitting down, maintaining balance or walking in a coordinated manner [3]. In this case, we refer to a qualitative evaluation of human movement, which passes through concepts such as ability (or motor skill) and takes into account factors linked to proprioception and intramuscular and intermuscular coordination [20]. Physical capacity is a quantitative measurement of the ability to carry out tasks without undue fatigue, reflecting the health of various physiological and neurological pathways. The most widely used capacity indicator is VO2max, with which the amount of oxygen used while exercising is measured, but other indicators are also used, such as speed and endurance [21].

### 2.1 Physical capacity and function tests

A plethora of specialised tests are adopted for objectively measuring physical capacity and function [22], such as the Cardiopulmonary Exercise Test (CPET) to measure maximal aerobic capacity (VO2max) on a treadmill or cycle ergometer, or the Fullerton Functional Fitness Test Battery [23, 24], which provides useful information on function in daily activities. Such tests tend to be lengthy, cumbersome to execute and require specialised personnel as well as costly equipment. Since our research focuses on the representativeness of data from free-living situations, CPET tests are not well-suited as points of comparison. Instead, clinical tests that are administered in a way which

shares more similarities with daily life are needed. This includes common activities such as walking, getting up, sitting down, and walking on stairs. Below, we elaborate on four groups of such clinical tests. We intentionally did not consider clinical scales such as the MDS-UPDRS scale for Parkinson's Disease, given their subjectivity and intra-rater variability [25].

The first group includes various distance- and time-based walking assessments. Examples of distance-based tests are the 10-meter walk test and the 400-meter walk test, which measure the time taken for the patient to cover a given distance and uses this as a measure of physical capacity [26, 27, 28]. An example of the time-based test is the 6MWT, often used in the monitoring of conditions such as Multiple Sclerosis (MS), pulmonary arterial hypertension (PAH) and other chronic diseases, as well as physical capacity in elderly populations [29, 30, 31]. Gait speed and endurance indicators from these tests are strongly correlated with the level of independence and life expectancy[32]. In terms of convenience and simplicity, these sub-maximal capacity tests do not require expensive instruments, but a stopwatch and a path with a known distance to walk are enough. Executing the tests is also a simple matter of walking for a certain distance or duration, which causes much less strain on test subjects whose physical condition often means they would struggle to complete maximal capacity tests. These aspects have made them widely adopted in gait analysis and capacity testing, especially for people with pathologies [33].

The second group focuses on the TUG test, which is of particular interest for neurological disorders such as Parkinson's [34]. The test asks the patient to stand up from a chair, walk straight for 3 meters, turn around, walk back and sit on the chair again. This procedure involves inclinations and rotations of the trunk axis, which can determine the functional qualities of the patient, including their balance. To collect data on these aspects, technologies such as sensorized mats and Inertial Measurements Units (IMUs) can be used, obtaining valuable information regarding the quality of movements such as gait symmetry, baropodometric load distribution and step rhythm [35, 36].

By only measuring performance time, the TUG test remains a strong indicator of the risk of falling or gait anomalies [35]. Through the use of a stopwatch, it is possible to establish the speed of execution of the test: this data point is able to reveal functional insufficiencies linked to pathologies. In this case, it is also a quick test to conduct, and, as with the walking tests, there is no requirement for expensive or bulky equipment to collect the performance time [37].

The third group of tests covers the Sit-to-Stand (STS) test, which is commonly performed among Chronic obstructive pulmonary disease (COPD) patients [38] for instance, or to estimate fall risk in PD patients [39]. The sit-to-stand action corresponds to the transition that a person performs from sitting to standing. The STS test is usually performed by doing multiple sit-to-stand transitions to evaluate balance and muscle strength [40]. In this category, the most adopted tests are the 5 times Sit-To-Stand Test (5xSTS) and the 30 seconds Sit-to-Stand Test (30STS). From these tests, one can consider the duration required to complete a given number of transitions or the number of repetitions within a selected time window, where an increased number of STS repetitions may also demonstrate muscle endurance capacity, for example, in the 30STS. To capture the outcomes of these tests it is required to have a stopwatch and a standard armless chair, in addition to the guidance and observation of experts.

Lastly, the fourth group of tests includes step tests. This type of testing is usually performed within occupational health and fitness evaluation [41], targeting healthy adults but also people with chronic heart diseases or COPD [42]. There are many types of step tests, which are distinguished by characteristics such as duration, step height, or the presence of phases such as initial warm-ups or heart rate thresholds to be reached for the test to end. For instance, in the specific case of the Chester Step Test, participants are required to step up and down a 30cm high platform at a progressively increasing rate set by a metronome or music beat. During this activity, the heart rate is monitored every two minutes; the overall test duration may vary from subject to subject.

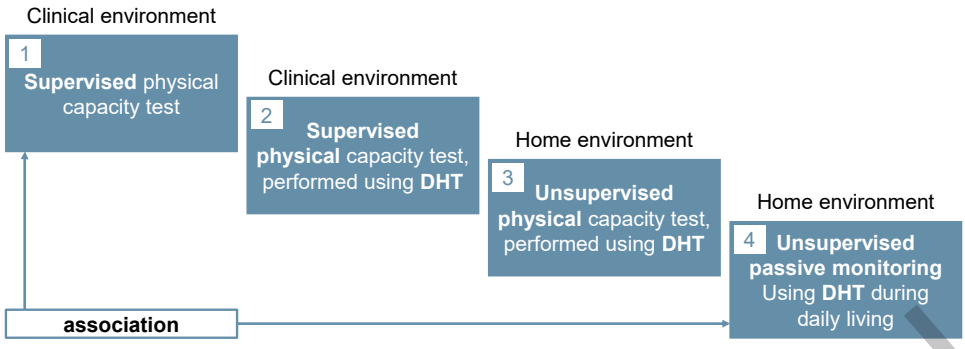


Fig. 1. Progression towards the introduction and use of technology in relation to physical assessments.

## 2.2 Physical capacity and function in free-living activities

Nowadays, DHT such as sensors in smartphones and wearables can capture many aspects of daily life activity, such as walking intervals, activity type, or heart rate, characterising a person's overall physical activity. The International Classification of Functioning, Disability and Health (ICF) [43] provides a framework for the characterisation of health, introducing the concept of physical activity as *performance*. This describes what an individual does in his or her environment, stressing the fact that this performance is dependent on the natural context of the individual. At the basis of our research is the hypothesis that information collected by DHTs in a natural environment, is impacted by the physical function and/or capacity of the people using the technology. DHT may be introduced in different contexts, including clinics, homes, and the community. In this way, clinicians would be provided with complementary information about performance [44]. In order to investigate the role of technology in relation to functional and capacity assessment, it can be helpful to view its introduction as a progression through four steps (Figure 1).

The first step corresponds to the performance of the standard physical test, performed in a clinical environment, with supervision and guidelines provided by experts. This is the traditional way of doing tests, and the use of sensor technology is usually absent. In the second step, DHT is introduced. The test is still performed in the clinic with an expert supervising the patient as well as operating any technology aids that tests may require to complement subjective expert observations. The use of a controlled environment is a common approach when developing methods and algorithms for DHT to ensure clinical relevance and accuracy. Step three involves patients reproducing the same test without the supervision of a clinical expert, making use of the technology. This method ensures that the patient understands and follows the expected testing process, mimicking a clinical environment but in their home or community environment. This approach has been extensively researched and validated, with a significant body of literature exploring its effectiveness and reliability [45, 46].

Building on these foundations, the fourth step involves using DHTs without performing controlled tests. This approach relies on data collected passively from participants as they focus on their everyday life activities, while the DHT assesses if this data can be associated with any clinically relevant tests. In other words, this approach aims to objectively measure performance and investigate how it relates to measures of capacity through data gathered in free-living conditions during their natural, daily activities [47]. While this method is still emerging, it represents a more seamless and unobtrusive means of monitoring, which is likely to allow participants to feel less like patients and more like healthy individuals. The collection of data within free-living conditions has the potential to provide richer insights into a patient's actual physical state during their everyday life, rather than the time-bound in-clinic tests on a given day. Such tests may, to a varying degree, be affected by external factors that only apply to that specific day, which introduces misleading test results despite the controlled environment

and procedure they follow. Despite its recognized potential [15], the validity and clinical actionability of data collected during daily living are not well understood.

## 3 Literature Review Methodology

### 3.1 Search Strategy

The search strategy used in this article follows the guidelines promoted by PRISMA [48]. Three main databases were considered in the identification phase for the literature review: Scopus, PubMed, and Web of Science. The literature search focused on articles published within the last decade (2014 to 2024), which were accessible and available in English. Papers from conference proceedings and journals were both considered. In the first screening phase, articles were distributed among five researchers and titles and abstracts were matched against inclusion and exclusion criteria. In a second eligibility phase, the full text of each article was assessed by the researchers. For articles deemed eligible, relevant information and key aspects were extracted and organized into tables for further analysis.

The queries used in the database were run on the 10th February 2025, and they are as follows:

(1) TS=("meter? walk* test" OR "minute? walk* test" OR ?mwt) AND TS=("free-living" OR "free living" OR "ecologic* monitor*" OR "passive* data collect*" OR "in-the-wild" OR "real-world" OR (remote* NEAR/3 collect*) OR "remotely collect*") AND TS=(sensor* OR mobile* OR smartphone* OR wearable* OR accelerometer OR device* OR smart* OR tracker) AND PY=(2014-2024)
(2) TS=("time* up and go" OR TUG) AND TS=("free-living" OR "free living" OR "ecologic* monitor*" OR "passive* data collect*" OR "in-the-wild" OR "real-world" OR (remote* NEAR/3 collect*) OR "remotely collect*") AND TS=(sensor* OR mobile* OR smartphone* OR wearable* OR accelerometer OR device* OR smart* OR tracker) AND PY=(2014-2024)
(3) TS=("sit to stand" OR STS OR "sit-to-stand") AND TS=("free-living" OR "free living" OR "ecologic* monitor*" OR "passive* data collect*" OR "in-the-wild" OR "real-world" OR (remote* NEAR/3 collect*) OR "remotely collect*") AND TS=(sensor* OR mobile* OR smartphone* OR wearable* OR accelerometer OR device* OR smart* OR tracker) AND PY=(2014-2024)
(4) TS=("step test" OR "chester step test" OR chester) AND TS=("free-living" OR "free living" OR "ecologic* monitor*" OR "passive* data collect*" OR "in-the-wild" OR "real-world" OR (remote* NEAR/3 collect*) OR "remotely collect*") AND TS=(sensor* OR mobile* OR smartphone* OR wearable* OR accelerometer OR device* OR smart* OR tracker) AND PY=(2014-2024)
(5) TS=("physical capacity" OR "physical function") AND TS=("free-living" OR "free living" OR "ecologic* monitor*" OR "passive* data collect*" OR "in-the-wild" OR "real-world" OR (remote* NEAR/3 collect*) OR "remotely collect*") AND TS=(sensor* OR mobile* OR smartphone* OR wearable* OR accelerometer OR device* OR smart* OR tracker) AND PY=(2014-2024)

The query structure is defined by three main components that were all considered relevant, and thus the AND operator is used between them. The first component is the type of physical test considered. As reported in Section 2, we consider four main groups of tests: walk tests, timed up and go test, sit to stand and step tests. The second component of the query corresponds to the data collection environment, which, in this case, is limited to free-living conditions. The third component pertains to the presence of technology, including wearable devices, smartphones, or sensors. While the data collection environment and technological aspects remain constant, the type of test varies to cover different groups. In addition to these test groups, we run a more general query in relation to "physical capacity" and "physical function". Given the relevance of physical function and capacity in this work, adding this query allows capturing broader range of articles in addition to the ones focusing on a specific test type.

### 3.2 Inclusion and exclusion criteria

Studies were included or excluded based on several criteria to ensure relevance and focus on the research objectives. Inclusion criteria included studies that (1) considered real-world or free-living data collection, (2) included one or more standard physical tests, and (3) performed an analysis between the real-world passively collected data and physical test outcomes (e.g. distance for the 6MWT). Reviews that aligned with the research topic were retained for related works. During the identification phase, duplicate records were removed. Articles were excluded if they were (1) out of the scope of the review, (2) only focused on reproducing physical tests in non-clinical environments, (3) did not make use of technology or sensors, or (4) the manuscript type was a protocol, editorial, or review.

### 3.3 Thematic analysis

In the field of remote activity monitoring in free-living conditions, particularly concerning assessments of physical function and capacity, we identified several key themes to explore within the reviewed articles. These aspects relate to the research questions outlined in Section 1.

To investigate the possibility of relating passively collected data to standard clinical tests (RQ1), we extracted the proposed associations between these two entities from each reviewed article, such as correlations between physical activity indicators and test results. From a more technical perspective, we extracted information (model, position, sensor modality) of the DHT used within each study (RQ2), and the metrics computed on the raw data collected in a free-living environment and how these are associated to the standard tests (RQ3). This is particularly helpful for future studies that aim to validate certain technological choices in determined contexts.

In relation to the experimental protocol (RQ4), we extracted details on the structure of study designs, including the physical tests adopted and the methods or instruments used to establish the ground truth for the physical assessments. As free-living monitoring is of particular interest, we analysed the monitoring duration window chosen in the articles to understand what is the most common duration required to collect physical activity data that can match clinical standards. We furthermore investigated the device-wearing time constraints proposed by the articles, which reflect the minimum conditions required to validate data collected in an unsupervised environment. Given the relevance of wearing time and participant collaboration in this type of study protocol, we finally reviewed the percentage of participants who adhered to the protocol outlined in the included papers. Adherence is defined here as the ratio between the number of participants who successfully completed the protocol respecting the minimum recommended duration and the number of recruited participants.

### 3.4 Risk of bias assessment

Research studies investigating DHT within cohort target populations may incur into biases of different types. We followed the methods suggested by Cochrane, particularly the ROBINS-I assessment, more tied to interventions, and the ROBINS-E assessment, tied to exposure, both used in non-randomized studies [49]. The articles included in this review are mostly cross-sectional observational studies, not focusing directly on interventional aspects; therefore, the bias assessment addresses two main types tied to pre-intervention stages, as suggested in Table 25.6.a from Sterne et al. [49], namely, confounding bias and selection bias. To obtain an overall risk of bias, we select the one with highest risk among the two.

For the confounding bias, we investigated whether the reporting of baseline physical fitness levels and other confounding factors is reported. Confounding bias is rated using the ROBINS scale as follows. Low risk: when relevant confounders are adequately measured and controlled for; Moderate risk: when some confounders are not fully accounted for but are unlikely to substantially affect outcomes; Serious risk: when important confounders are missing, poorly measured, or not adjusted for; Critical risk: when confounding is so severe that the study estimates are likely to be substantially distorted.

For the selection bias, we examined how and how many participants were identified and recruited, and whether the included sample differed systematically from the target population. Selection bias is likewise rated using the ROBINS scale. Low risk: when recruitment is clearly described, sample size is greater than 15 and the group is representative of the target population; Moderate risk: when sample size is greater than 15, and minor deviations from representativeness are present; Serious risk: when sample size is less than 16, or when recruitment or exclusions result in a sample that may differ systematically from the target population; Critical risk: when sample size is less than 16 or when the selection process severely undermines the validity or generalisability of the results.

To address this aspect of the systematic literature review, we made use of the AI-based tool Anara [50] to highlight sections of the papers under review related to 2 concrete questions: 1) Does the article present confounding bias? and 2) does the article present selection bias?. All outputs provided by Anara were further manually verified by one researcher vis-a-vis the text of the paper. This ensured efficiency and robustness to the investigation. The Anara tool was used only for risk of bias assessment, and no other parts of the review were addressed through this approach.

## 4 Results

The following section presents the findings of the systematic review and begins with an overview of the search results, followed by a description of the populations and health conditions addressed in the articles. The remainder of the section presents the themes that emerged during our thematic analysis of the identified articles. This includes an outline of DHT, an analysis of the experimental protocols used, and a report on extracted metrics and their relationships with standard clinical assessments. The section concludes by reporting on wearing time and user adherence from the identified articles.

### 4.1 Search results and risk of bias assessment

A total of 678 articles were initially identified from the three databases, Scopus, PubMed, and Web of Science, with the defined search criteria described in Section 3. Of these, 39 studies passed all inclusion criteria and are assessed in this review. In addition, 8 studies were added through snowballing techniques, such as by reviewing relevant cited references, obtaining a total of 47 articles. Figure 2 shows the flowchart corresponding to the phases of the search, screening and review process, while Table 1 shows the initial results from the query for each category. After removing duplicate articles across databases, 298 were left for investigation. To track the progression of research over time, we examined the publication year of the studies. Figure 3 illustrates the distribution of publication years across the selected papers. The distribution shows articles from 2009 until today, highlighting the increasing relevance of this research topic over time.

Table 1. Query search results and number of articles included in the review.

| | **Scopus** | **Pubmed** | **Web of Science** | **Total** |
|---|---|---|---|---|
| MWT query | 22 | 43 | 52 | 117 |
| TUG query | 37 | 32 | 31 | 100 |
| STS query | 98 | 76 | 86 | 260 |
| STEP query | 7 | 6 | 7 | 20 |
| CAPACITY query | 70 | 58 | 53 | 181 |
| | | | | 678 |

Confounding and selection bias were assessed across the articles, resulting in 10/47 articles with an overall low risk of bias, 31/47 with a moderate risk of bias, 6/47 with a serious risk of bias and no articles with a critical risk of

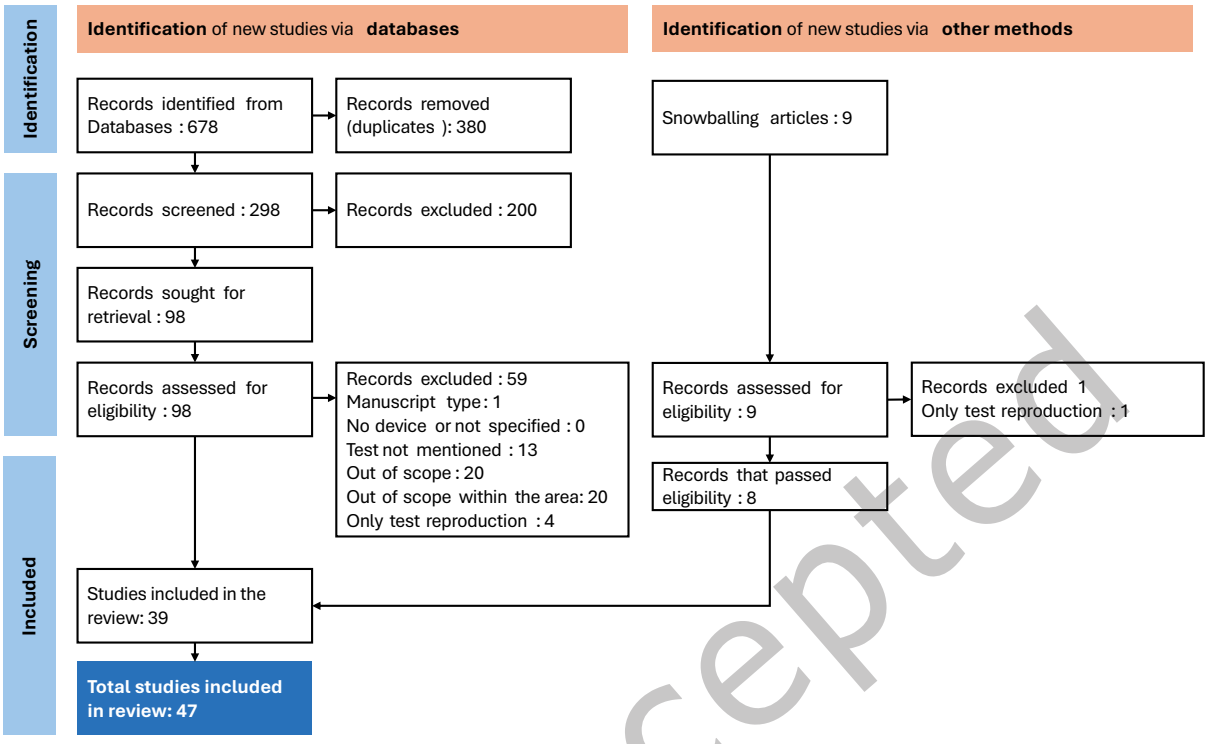


Fig. 2. Systematic review flowchart according to the different stages. Following PRISMA guidelines [48].

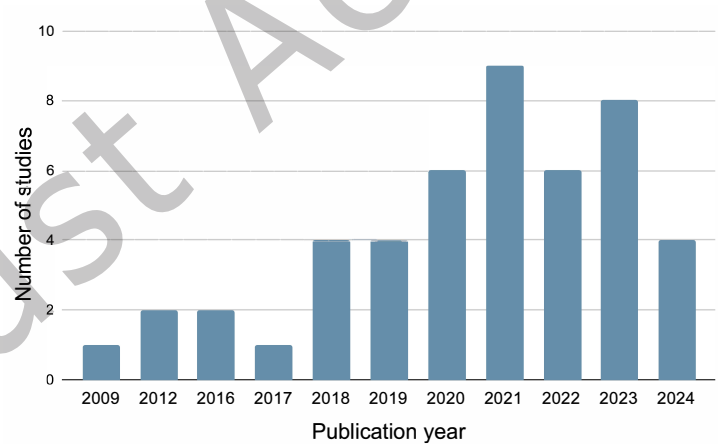


Fig. 3. Publication year distribution across the studies selected for the review.

bias. In particular, articles with a serious risk of bias included those with 15 or less number of analysed participants ([51, 52, 53, 54]); the white-paper from Apple Watch [55] that considered participants non-representative of the U.S. population with COPD, and, lastly, the work from Sokas et al. [56] which considered as baseline physical

assessment the outcome of standardised equations including age, weight, sex, and maximal heart rate. Table 8, in the appendix, shows the outcome of the risk of bias assessment for each article.

## 4.2 Populations

The populations included in this review represent a diverse range of participants, for a total number of 7910 subjects across all studies and with various health conditions. As visible in Table 2, out of 47 articles, 14 studies (30%) did not explore explicit conditions and instead include common-dwelling adults or elderly [57, 58, 59, 60, 55, 61, 62, 63, 64, 65, 66, 67, 68, 69], 9 studies investigate cases of MS (19%) [70, 71, 72, 73, 74, 75, 76, 77, 78], 6 (13%) studies consider cardiovascular conditions [79, 56, 53, 52, 80, 81], and 4 (8.5%) consider Parkinson's disease [82, 8, 83, 84]. We group musculoskeletal conditions as juvenile idiopathic arthritis, knee osteoarthritis or frailty [85, 86, 87]. Neurological conditions as stroke, or cerebral palsy [54, 88, 89], and at last neuromuscular disorders such as fascioscapulohumeral dystrophy, autoimmune myasthenia gravis or Duchenne muscular dystrophy [90, 91, 92]. Other conditions or contexts are observed in single articles, such as chronic kidney disease [93], pre-elective surgery [94], or fallers [51].

The population sample sizes are between 12 [82] and 2001 participants [86].

Table 2. Conditions reported in the 47 included articles.

| **Condition** | |
|---|---|
| Multiple Sclerosis | 9 |
| Cardiovascular Disease | 6 |
| Parkinson's | 4 |
| Musculoskeletal Disease | 3 |
| Neurological Condition | 3 |
| Neuromuscolar Disorder | 3 |
| Cancer (survivors) | 2 |
| After Orthopedic Surgery | 1 |
| Chronic Kidney Disease | 1 |
| Fallers | 1 |
| Pre-elective Surgery | 1 |
| Unspecified condition | 13 |

## 4.3 Digital Health Technologies

The articles analysed in this literature review use a range of wearable technologies and sensors, with a strong focus on IMUs, used mainly for raw accelerometry, physical activity type, and step count. IMUs can be embedded in smartphones or in wearable devices, like smartwatches and fitness trackers (e.g. Fitbit, Withings Steel HR smartwatch, Apple Watch, etc.), and other wrist-worn actigraphy devices (e.g. GENEActiv, ActiGraph, ActivPal, etc.). Other studies incorporate Global Positioning System (GPS) for location tracking for spatial motion analysis. Within the reviewed studies, only one does not use technology directly attached to the participant's body. Instead, [59] utilises a system of ambient sensors (presence sensors, door and fridge sensors, and bed sensors) to monitor the individual's activities within their home environment. Figure 4 illustrates different body positions, sensor modalities and device models included in the articles.

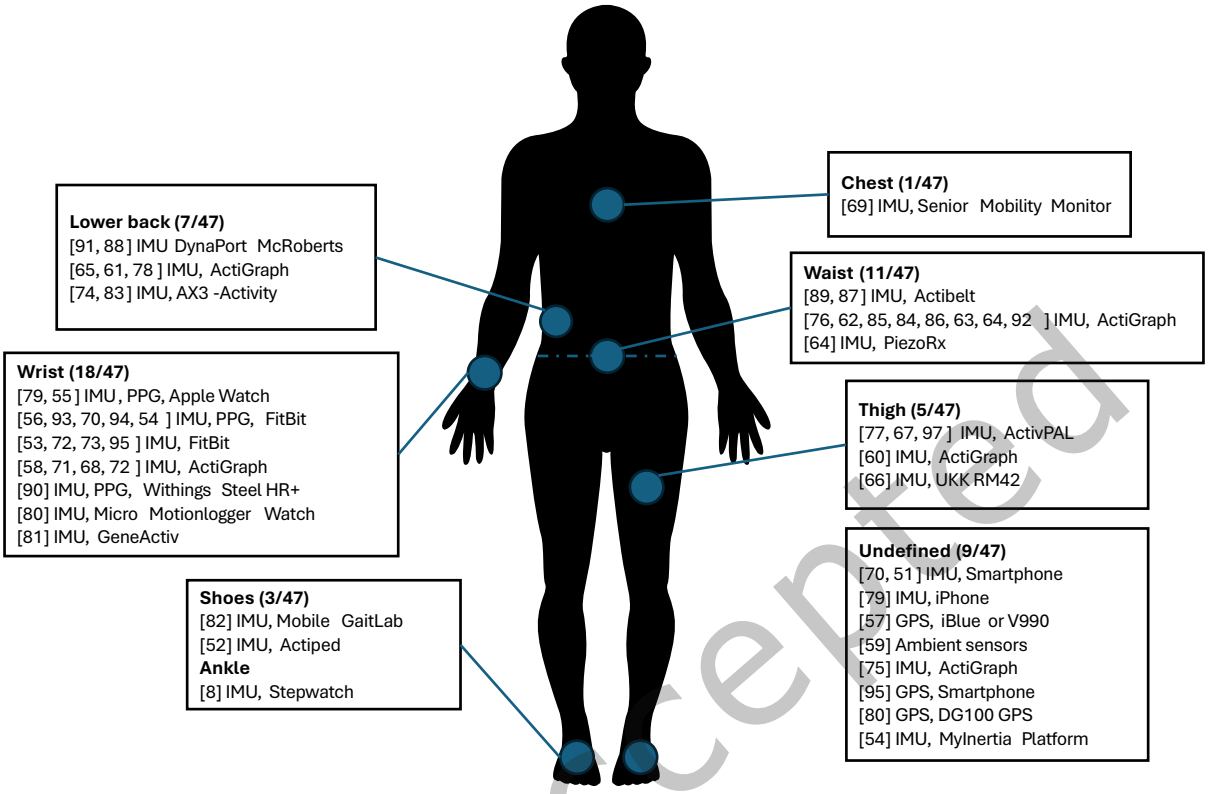


Fig. 4. Human body representation with highlights on device position, sensor modality and brand.



## 4.4 Experimental protocol

The experimental protocols of the included papers are generally based on cross-sectional designs, with an initial supervised visit, typically conducted in a clinic in the presence of expert researchers or clinical staff, followed by a period of real-world data collection during which participants wear or carry a device with them. Supervised visits typically involve mobility assessments, during which patients perform physical tests under the observation of expert clinicians or researchers. In some cases, experts also employ tools to establish reference outcomes for the tests. For example, distance is often measured using trundle wheels [89, 87, 93, 53, 74, 75, 78], ergometer [85], pressured gait mats [57, 84], photocells [88] or using floor markings at known distances. Time duration is commonly recorded using a stopwatch [53]. To accurately count steps, additional inertial sensors are used to establish ground truth step counts [58, 72, 57]. If no instrumentation is specified, standard clinical guidelines are assumed to be used for running supervised physical tests. Only two studies did not include an initial visit. The first one [82] asked two trained human annotators to label the start and stop of a test performed by the participant in their home environment, while in the second case [56], the 6MWD was estimated using an equation that uses anthropometric variables.

Table 3 summarises the clinical tests across the 47 articles. In particular, some less common tests that shared certain characteristics were grouped together:

- X-Meter Walk Tests: Walking tests based on fixed lengths such as 3m, 4m, 10m, 20m, or 400m.

- X-Minute Walk Tests: Walking tests based on fixed duration such as 1, 2, 4, 10 minute walk tests.
- STS related Tests: 5xSTS, STS, 30STS.
- Walking Capacity Tests: Shuttle test, Balke Treadmill Test, Gardner-Skinner treadmill protocol, and the Strandness treadmill protocol.
- Balance Tests: Sway test, Mini-BESTest, standing balance test, Tinetti Performance Oriented Mobility Assessment (POMA)
- Lower Limb Capacity Tests: Chair sit-and-reach, Watt-maxtest, Isometric Knee Extension strength
- Upper Limb Capacity Tests: Back scratch, arm curl

Table 3. Standard physical test assessment and their frequency in the reviewed studies.

| Test | #Articles |
|---|---|
| 6 Minute Walk Test (6MWT) | 26 |
| X-Meter Walk Tests | 16 |
| Timed Up and Go Test (TUG) | 12 |
| X-Minute (other than 6) Walk Tests | 8 |
| Sit-to-Stand (STS) related Tests | 7 |
| Timed 25-Foot Walk Test (T25FW) | 6 |
| Walking Capacity Tests | 5 |
| Short Physical Performance Battery Test (SPPB) | 5 |
| Balance Tests | 4 |
| Hand Grip Strength Test | 4 |
| Lower Limb Capacity Tests | 3 |
| Upper Limb Capacity Tests | 2 |
| Figure of Eight Walk Test | 1 |
| Four Square Step Test (FSST) | 1 |



In addition to physical assessments, patients are sometimes required to complete questionnaires aimed at investigating various aspects of their health and daily life. These questionnaires may be the Health-related Quality of Life (HRQoL) questionnaire [53] to explore various dimensions of an individual's well-being, the Clinical and Patient Global Impression scales (CGI and PGI, respectively) to assess patients' global functioning [83], or the Expanded Disability Status Scale (EDSS) [72] which targets particularly MS patients. While we recognize the relevance of questionnaires to complement physical tests, these fall outside of the scope of this review.

After the first supervised visit, the experimental protocols rely on a period of real-world data collection. During this passive monitoring phase, participants may be asked to perform unsupervised physical tests as a complement to continuously wearing the included devices as they pursue their daily routines and activities. As shown in Figure 5, the most common monitoring duration is one week, even though longer periods of time are also considered, such as 6 months [79], between 6 and 24 months [59, 72, 70] and an unspecified period of time [55]. In the latter study, even though the monitoring duration is not specified, the data processing phase for algorithm development is likely to be referred to a time window selected by researchers.

## 4.5 Metrics of passive monitoring and relationship with physical tests

Table 4 reports the extracted summary metrics from the reviewed studies. The most frequent metric is related to step count and multiple statistics of this variable, such as average per day, or standard deviation of daily steps. The second most frequent metric corresponds to the time spent across different activity intensity types. The third

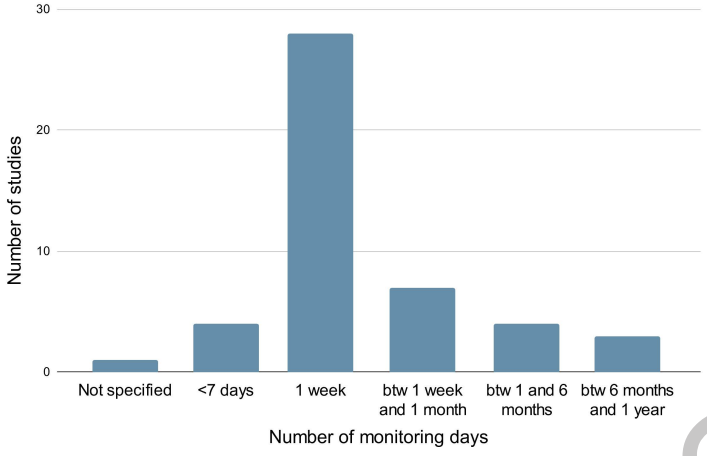


Fig. 5. Free-living monitoring duration across the articles.

most frequent type of metrics are gait parameters extracted from accelerometry data, such as walking symmetry and stride-to-stride pattern [69, 51]

Table 4. Categorized summary metrics across the articles.

| **Summary metrics** | **#Articles** |
|---|---|
| Steps | 23 |
| Sedentary-light-MVPA activities | 18 |
| Gait parameters (from IMU) | 11 |
| Raw accelerometry statistics | 10 |
| Movement intensity | 7 |
| Speed | 5 |
| Localization parameters | 4 |
| Heart rate | 3 |
| Sit to stand parameters | 3 |
| Postural transition | 2 |
| Ambient sensors parameters | 1 |
| Stepping endurance [duration] | 1 |
| Heart rate over steps | 1 |
| Smartphone interactions | 1 |
| Wear time | 1 |
| Walked distance | 1 |

Table 7, in the Appendix A, reports, for each article, the association between passively recorded data and one or more standard tests among the ones in Table 3. Out of 47, 37 articles (78.7%) do correlation analysis or provide more general considerations on the relation between passively collected data and standard tests. In addition, 7/47 (14.9%) articles use passive data to predict the outcomes of the 6MWT or the TUG, classifying classes of performance, while 2/47 articles (4.3%) provide a regression estimation of the TUG duration.

Figure 7 and Figure 6 show indicators of the association between summary metrics and the test of interest. For the 6MWT, step metrics are used in a multitude of articles reporting significant correlation values between 0.21 and 0.68. The metrics with the highest correlation with the 6MWT are the NET-F index (0.68, $p < 0.001$) [94] in a population of pre-elective surgery patients, the Peak-30CAD and Peak-1CAD for the MS population (0.68, $p < 0.01$ and 0.67, $p < 0.001$) [76], and, for MS patients, the average daily movement (0.63,$p < 0.001$) [75]. For the TUG, the data aggregation techniques that best align with the clinical assessment are the empirical TUG estimation from Silva et al. [51] (correlation of 0.89) and the estimated TUG value from Saporito et al. [69] (0.7 correlation and Area Under the Curve (AUC) of 0.89). The estimate from [51] is obtained from a regression model whose features include demographic information, gait speed and stride, number of steps, a postural transition parameter and the result of a questionnaire on fear of falling. Notable in relation to the TUG test are the results from Low et al. [95], which show that TUG duration correlates significantly with peak gait cadence (-0.72) and daily step count (-0.61) but not with localization information, possibly because it is unsuitable and too diverse. The estimate from [69] relies on six mobility indicators: walking quantity and quality, chair-rise ability, and durations of active and inactive bouts.

While most works focused on the regression of the test outcomes, some others provide a classification of the test outcome in predefined classes, such as above or below a certain threshold. Table 5 shows the values of roc-AUC obtained for the classification of the test results. In particular, for the 6MWT, Rens et al. [79] binary classify the 6MWD with a threshold of 300-meter, while [55] uses a threshold of 360-meter. Sun et al. [70] consider the upper 25% and lower 25% of the reference values. Regarding the TUG duration, [59] uses a cut-off threshold of 12-second, while [69] sets a 10-second threshold.

Table 5. AUC values for the articles that perform classification of the 6MWT and TUG test outcomes.

| **6MWD AUC** | | **TUG AUC** | |
|---|---|---|---|
| 300m threshold [79] | 0.64 | 12s threshold [59] | 0.79 |
| 360m threshold [55] | 0.95 | 10s threshold [69] | 0.89 |
| upper and lower 25% 6MWD [70] | 0.87 | | |

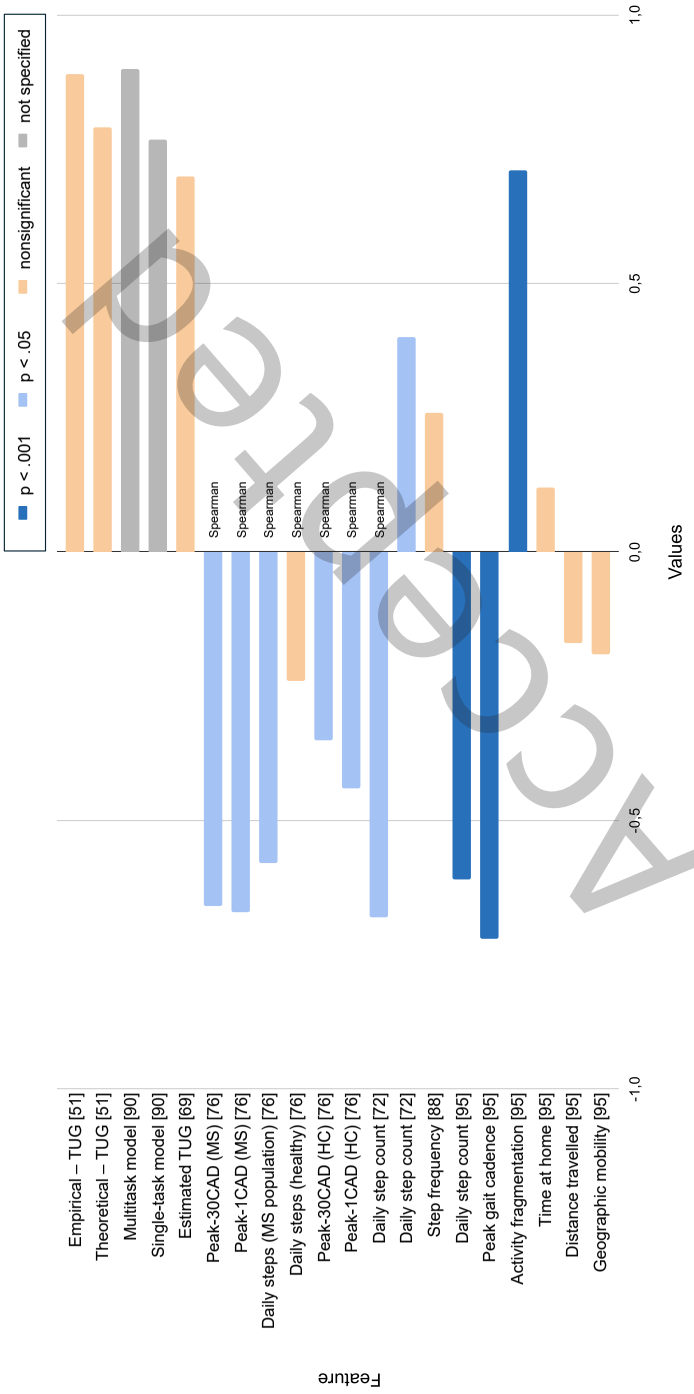


Fig. 6. Association between free-living features and the TUG duration.
HC = Healthy Control; CAD = Cadence; MS = Multiple Sclerosis.

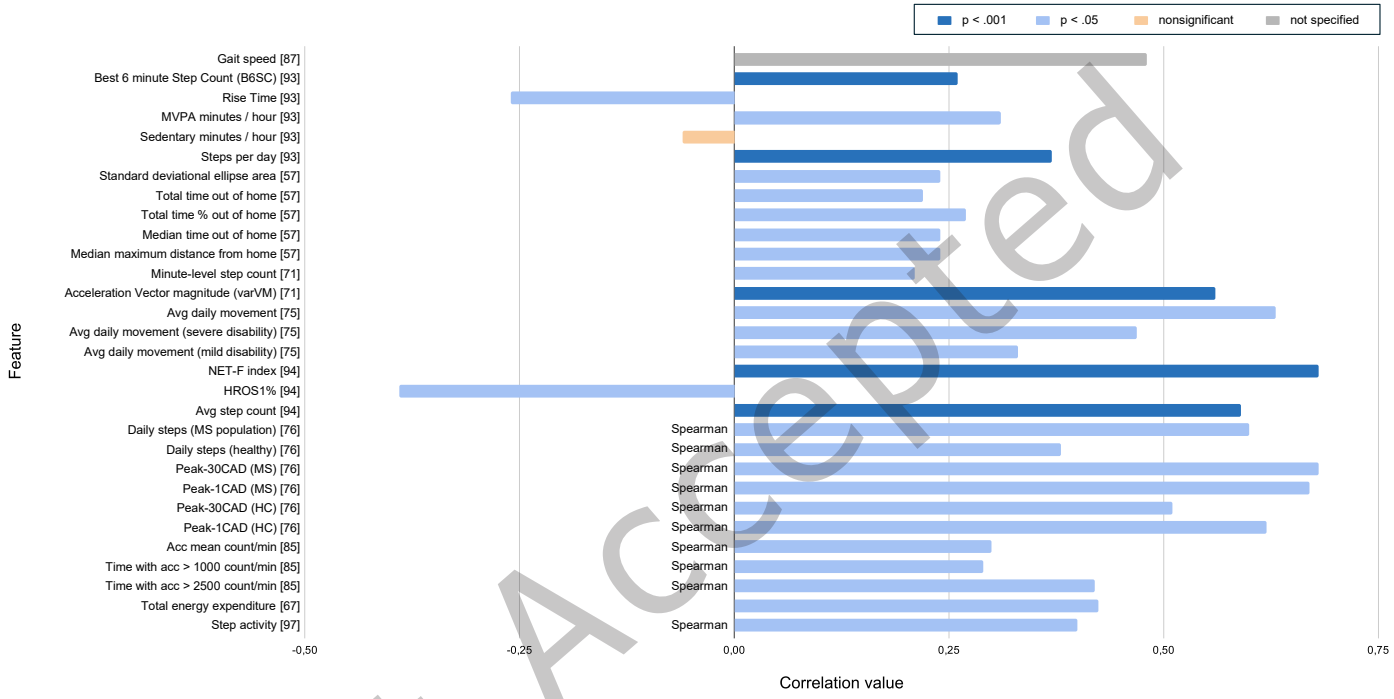


Fig. 7. Correlation between free-living features and the 6MWT distance.
HROS = Heart Rate Over Steps; NET-F = Non-Exercise Testing Cardiorespiratory Fitness; CAD = Cadence; MVPA = Moderate to Vigorous Physical Activity.

Table 6. Minimum amount of data for collecting a valid day and valid collection (more days). PIR = Passive infrared sensor.

| **Valid day** | hours >3, 8, 9, 10, 14, or 18 | steps >0, 100, 128 or 300 | PIR >300s | Walk >5 min | |
|---|---|---|---|---|---|
| **Articles** | [69, 55, 85, 97, 76, 93, 75] [62, 86, 64, 89, 53, 92, 95] [80, 84, 87, 65, 68] | [72, 79, 70, 54] | [59] | [88] | |
| **Valid collection** | 30/84 or 10/28 days | 2, 3, 4, or 6 days out of 7 | At least 1 day | 5 or 10 /14 days | 3/4 days |
| **Articles** | [69, 55] | [60, 75, 85, 87, 66, 81, 93] [84, 64, 53, 91, 92, 80, 61] [57, 95, 55] | [76] | [90, 59] | [62] |

### 4.6 Wearing time and user adherence

Device wearing time is essential when relying on passive data collection from wearable devices. The majority of the articles placed significant emphasis on this aspect, enforcing specific constraints for including or excluding data, while others acknowledged that they were missing some data or had non-wear time periods Some studies investigated methods for detecting non-wear time. For example, [71] uses the method from Choi et al. [96], which considers intervals of 90 minutes of activity as absence, while [69] classifies non-wear time as the absence of movement intervals within 15 minutes, which is similar in approach to [62] who consider a window of 60 minutes with no movement as absence.

The majority of articles report considerations on the minimum valid amount of data when doing real-world data analysis. This includes constraints for considering a day as having sufficient data or longer time periods. As seen in Table 6, most works base the minimum threshold of collected data on duration restrictions, while others focus on sufficient sensor information (e.g. steps >threshold for the step count to be valid on a single day).

Finally, Figure 8 reports the adherence percentage, monitoring duration and population size of each study, together with a simplified visualization of the same information. The ratio [days/days] corresponds to the number of days the user adhered to the protocol divided by the expected overall number of days, whereas the percentage without notation corresponds to the number of people who successfully completed the protocol versus the overall number of people included in the studies. Studies that consider a monitoring duration of one week have an average adherence of 82%. Some studies report adherence not as the number of compliant participants among the recruited participants, but as the valid collected days over the expected number of worn days.

## 5 Discussion

This systematic review analyses 47 articles related to the feasibility of associating free-living DHT data to physical tests. It aims to provide a comprehensive overview of the current state of research on the relationship between passive monitoring and standard clinical assessment and to identify and analyse the most common methods and practices within these studies. After presenting our reflections on the PRISMA [48] based search strategy, we discuss our findings in relation to each of our four research questions, where RQ1 is more general while RQ2-4 break down our results from the reviewed papers down into more nuance, including suggestions for further research associated with each RQ.

### 5.1 Search results and risk of bias assessment

Five queries were run to include a wide variety of physical function and capacity tests that use movement patterns that are commonly performed during everyday living, such as walking, transitioning from sitting to standing, or ascending stairs. We experimented with the query engines of Scopus, PubMed, and Web of Science databases. The fifth query was included to broaden the results beyond the specific tests we had identified as relevant for

| Article | User adherence | Monitoring days | Population size |
|---|---|---|---|
| [77] Mate et al. | 99% [days/days] | 7 | 125 |
| [75] Motl et al. | 98% | 7 | 256 |
| [69] Saporito et al. | 94% | 3 | 254 |
| [84] Nero et al. | 91% | 7 | 100 |
| [92] Nair et al. (*1w*) | 90% | 7 | 127 |
| [68] Choudhury et al. | 90% | 7 | 94 |
| [66] Karavirta et al. | 90% | 7 | 495 |
| [85] Nørgaard et al. | 87% | 7 | 70 |
| [71] Zhai et al. | 87% | 7 | 158 |
| [91] Birnbaum et al. | 87% | 7 | 38 |
| [88] van de Port et al. | 86% | 7 | 47 |
| [65] van Gameren et al. | 85% | 7 | 229 |
| [95] Low et al. | 84% [days/days] | 30 | 40 |
| [72] Block et al. | 83% [retention] | 365 | 63 |
| [94] Angelucci et al. | 83% | 7 | 29 |
| [57] Crane et al. | 82% | 7 | 182 |
| [81] Rowlands et al. | 82% | 7 | 238 |
| [53] Straiton et al. | 73% [days/days] | 8 | 15 |
| [80] Chaudru et al. | 73% | 7 | 23 |
| [79] Rens et al. | 73% | 210 | 107 |
| [61] Wright et al. | 72% | 7 | 105 |
| [56] Sokas et al. | 71% [days/days] | 7 | 99 |
| [89] Wiedmann et al. | 69% | 7 | 88 |
| [90] Zhuparris et al. | 66% [days/days] | 42 | 38 |
| [76] Zheng et al. | 63% | 7 | 321 |
| [59] Schütz et al. | 57% | 365 | 21 |
| [97] Bachman et al. | 57% | 7 | 86 |
| [54] Katzan et al. | 53% | 90 | 15 |
| [92] Nair et al. (*1yr*) | 45% | 365 | 127 |
| [83] Bouça-Machado et al. | 41% | 6 | 39 |
| [92] Nair et al. (*2yrs*) | 23% | 730 | 127 |

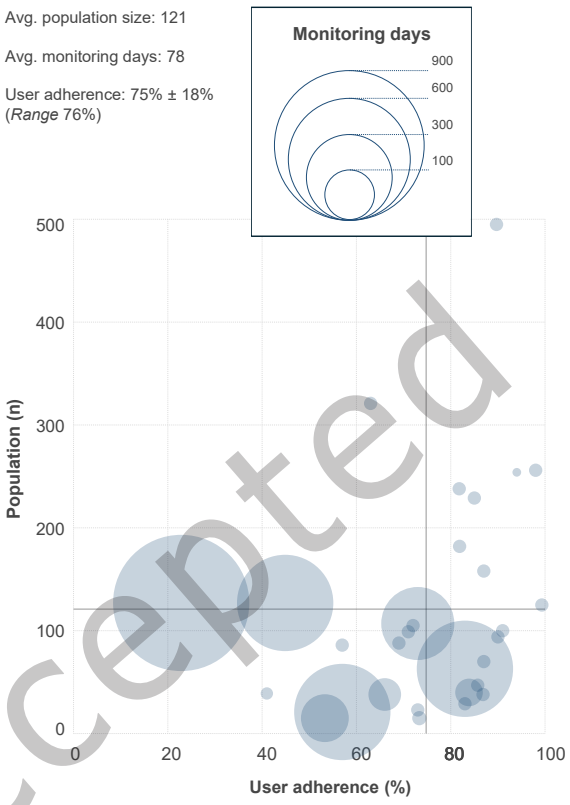


Fig. 8. Participants adherence in relation to the number of participants and monitoring duration. The table on the left side provides all measurements and articles are reported in descendent order based on user adherence. The three dimensional figure on the right side shows user adherence on the X-axis, population number on the Y-axis, and the size of the circle represents the number of monitoring days considered for the study. This is shown in the upper right legend. The grey lines correspond to the medians of each axis.



function and capacity. Our search yielded 39 articles that met our inclusion criteria. To expand the number of papers included in the review, we reviewed references included in the 39 studies that our queries had provided us with. This resulted in the addition of 8 studies, for a total of 47 papers. As a sanity check, we investigated why the papers obtained through snowballing techniques were not included in the query results. Three of those papers were not found through our queries because they were published before 2014. We set 2014 as a starting point since technology penetration has primarily escalated during the last decade [98] and further expect publications to follow the growing technology trend during the last and coming five years [99]. Four articles did not use keywords within the article title or abstract that allowed us to identify the physical tests they used or the type of technology they employed. Finally, one article is a white-paper from an industry source ([55]) and was therefore not indexed in the considered databases.

We investigated the risk of bias across the included manuscripts, in particular the confounding and selection biases. The analysis showed that the majority of studies (31/47) had a moderate risk of bias. This is largely due to the shared challenge of recruiting participants representative of specific health conditions and populations. Most

studies, in fact, include a relatively healthy population compared to the condition under investigation. This is most likely due to the challenges inherent in the use of technology, even passive sensors, when subjects with limited mobility are involved. It is natural, especially in a pilot study, to include those participants who are more likely to be willing and capable of using the technology throughout the duration of the study. This is also visible in the data where, with a few exceptions, adherence decreases as the number of monitoring days increases.

### 5.2 Associating passively collected data with physical tests (RQ1)

Most studies perform correlation analyses between passively collected data and standard clinical tests. While correlation can highlight the significant association between passive measurements and clinical test results, it does not clarify the underlying causal mechanisms driving these relationships, which can be important in determining applicability to clinical practice. Two studies establish physical capacity and function scores derived during the baseline assessment. Both show a positive correlation between the baseline score and a physical activity metric. In particular, Santos et al. [62] calculate a functional fitness score from a multitude of physical tests and observe a positive correlation with the duration of Moderate to Vigorous Physical Activity (MVPA) and a negative correlation with sedentary duration. Wright et al. [61] define physical activity as "what you do", while physical capacity as "what you can do", encapsulating multiple parameters within each term (e.g. active duration and movement intensity for physical activity, and 6MWT distance or STS duration for physical capacity). Based on this, they report a positive correlation between physical activity as performance and physical capacity.

Most articles show a significant correlation between free-living-derived metrics and standard physical assessments. However, a few articles report contrasting outcomes, claiming the absence of an association between metrics and the tests. In particular, Rekant et al. [58] note that "neither summary measures of global health nor physical activity were significantly related to 6MWT performance". This could be an effect of the considered population being veterans interested in health and wellness programs without any specific condition, capable of walking independently, or the fact that the monitoring duration for this group was only 1 day. Such a short duration may not be representative of the general physical activity of the participant, being influenced by multiple factors that may only occur on a single day. Moreover, Straiton et al. [53] report that physical activity in the free-living environment is not associated with physical capacity post-aortic valve replacement in clinical settings. In this case, the population is a sample of 15 patients with aortic stenosis, and the window of the investigation was four days before cardiovascular surgery and after (1 month). A reason for the absence of correlation may be the low sample size and the specific condition as well as reactions to surgery that can vary significantly from subject to subject. Following, van Gameren et al. [65] declare a weak correlation between the Short Physical Performance Battery (SPPB) test and daily life gait quality and quantity. The authors justify this outcome from the wide inter-individual variation of physical activity during daily life and the narrow variation of SPPB scores.

In two studies, sedentary behaviour was not associated with the 6MWD nor any physical fitness measure [91, 64], in contrast, sedentary time was correlated to STS performance in the work from Choudhury et al. [68] and it was negatively correlated (-0.32, $p<0.05$) to submaximal VO2 in [97]. Karle et al. [78] did not find correlations between the 2-Minute Walk Test (2MWT) cadence and the duration of walking time at 100%, 90%, 80%, and 70% of 2MWT cadence. They explain the lack of correlation by the fact that they wanted to compare higher intensity walking correspondence, differing from other works that focused on more generic metrics such as steps per day [72].

Seven studies ([72, 70, 59, 94, 69, 55]) introduce a model that classifies daily living physical activity into distinct performance classes rather than relying on regression analysis (Table 5). Stratifying physical activity in relation to a standard test into categories (e.g., high vs. low performance) may be simpler and more accurate from an algorithmic perspective. However, the significance of a binary outcome should be further investigated in relation to its applicability within a clinical context, especially in relation to decision-making. This remains an unexplored

area, but it is likely that this information could still be helpful as a complement to other assessments and should motivate further research in the direction of finer classifications through the use of time series analysis approaches that combine machine learning and other advanced techniques.

While the association between metrics obtained from free-living physical activity and standard assessment is significant in the majority of the proposed works (see Table 7), there is disagreement in some of the articles. This could be due to the focus on different populations and conditions. Participants with small mobility impairments likely show high variability in daily movement (the "what you do") with little association with capacity or function (the "what you can do"). This indicates a clear need for further investigation in this area, particularly to understand generalizability and differences across population demographics and health conditions.

### 5.3 Digital Health Technologies (RQ2)

Sensors and devices are the means that allow collecting data from participants' during their everyday life. Given how different daily activities and environments are, it is crucial to discuss DHT usage across the studies. This review identifies the most common device location as the wrist (18/47), largely as a result of the popularized wrist-worn devices from brands such as FitBit, Apple Watch, Withings, and ActiGraph. Following, other popular wear positions are the waist/hip position (11/47). This finding is in agreement with a review study from Olmedo et al. [12], on the use of wearable for elderly remote monitoring where it was shown that, among 56 studies, the most frequent devices are wrist-worn watches (25%), followed by wrist-worn bracelets (17%) and patches (17%). Another review by Storm et al., however, that focuses on assessing the 6MWT with wearable inertial sensors [100], reports that the most common device position is the lower back (18/28 articles), followed by the shank (8/28 articles) and ankle (7/28 articles). Compared with our findings, two differences can be highlighted: first, the focus only on the inertial sensing type of technology used and second, their focus on the reproduction of the test is in line with what we define as the 3rd step in Figure 1, rather than the 4th step where unsupervised collected data from everyday life is associated with physical tests. Regardless of the step targeted, findings are relevant to the advancement of the field. The more integrated DHT are expected to be in the everyday life of users, the more important it becomes to identify device positions that provide high-quality data as well as accommodate comfort and acceptance from users.

Furthermore, with a smartphone penetration rate higher than 85% globally [101], these devices are a natural DHT to include and/or to integrate wearables. Some differences come with relying on smartphone sensors over wearables, in part as a result of the body position, where a smartphone cannot be assumed to always be in the same location. Still, smartphones and wearables are both promising to collect data and insights about one's mobility.

In addition to IMUs and video sensory data, localization information is used in the study from Crane et al. [57], Zhuparris et al. [90], Santos et al. [55] and Low et al. [95]. In particular, studies consider solely space-related features such as the maximum distance from the home environment, time spent out-of-home, area of the distance covered outside and compactness [57, 95, 90]. While the industry white-paper [55] uses localization information only with the goal of estimating a more accurate 6MWD. The results from these studies are promising, but more investigation is needed to cover the privacy and ethical aspects concerning the continuous passive monitoring of one's location.

Somewhat surprisingly to us, no video cameras were present in the reviewed studies, even though this type of sensor is highly spread in multiple contexts (e.g., surveillance and gait analysis) and is natural to many smartphone users for documenting personal or social experiences. This may be due to several factors, ranging from recording of videos or pictures being viewed as potentially privacy-invasive to the fact that videos are computationally expensive to store and analyse or that cameras are often impractical and obtrusive to use in daily life contexts for

recording symptoms. Some of the included studies made minor use of cameras, where, e.g. [87] relies on video to capture steps as part of their ground truth measurements, but not as devices to be used to measure daily activity.

### 5.4 Metrics extraction and their association with physical tests (RQ3)

A key aspect of the reviewed articles is their emphasis on extracted metrics and their association with the outcomes of standard physical tests. These statistics vary significantly depending on the type of acquisition sensor, the observation window, and, most importantly, the nature of the feature itself and what it represents.

Statistics related to step count are the most common ones across the reviewed articles. This may be a result of steps being a good proxy for distance estimation and/or overall mobility throughout the day, which both carry potential meaning in terms of function and capacity. Examples of such features are the best 6-minute step count (B6SC) [93], which captures the highest number of steps performed during a window of six minutes throughout the day. This metric has a correlation of 0.26 ($p < 0.01$) with 6MWT. Angelucci et al. [94] consider the average step count with a correlation of 0.59 ($p < 0.01$). Steps per day obtain a correlation of 0.37 with 6MWT [93], and equally with STS in the work from O'Brien et al. [64].

Different studies end up with significantly different results on relatively similar features, indicating the need for finer investigation across groups and protocols. Very often used are also metrics related to the duration of sedentary-light-or moderate-to-vigorous activities. These show to correlate with a variety of physical tests such as the 6MWT [93, 66], STS [68], or with submaximal VO2 [97].

Among the metrics with the highest correlation with the 6MWT we find the Non-Exercise Testing Cardiorespiratory Fitness (NET-F index) (0.68, $p < 0.001$) [94], which corresponds to a combination of parameters such as the sex, age, BMI, resting HR, and physical activity levels, but also in relation to walking intensity, Peak-30CAD and Peak-1CAD for the MS population (0.68, $p < 0.01$ and 0.67, $p < 0.001$ [76]. These two metrics are determined by taking the average cadence of the highest 30 (or 1) non-consecutive minutes recorded in a day. Lastly, among highly correlated statistics, we find the average daily movement (0.63, $p < 0.001$) [75], derived from the vertical axes of the accelerometer. Studies also use time spent in activities of different intensities, such as [93], that correlates time spent in MVPA minutes every hour with the 6MWT obtaining 0.31 ($p < 0.01$); Santos et al. [62] report that sedentary behaviour is negatively associated with a fitness functional score, upper and lower body strength, balance and lower flexibility; and [71] which notes that there is a stronger association between MVPA and and clinical measures in comparison to steps per minute. Another reason why steps and activity intensity are frequently considered is their wide availability and accessibility from commercial devices such as wrist-worn devices and smartphones.

In the analysis from Zheng et al. [76] daily step count in a healthy control group and in a population suffering from MS show correlations of 0.38 and 0.60 $p < 0.01$ with 6MWD, respectively. Correlation values related to the MS population are higher in comparison to those of the healthy control group in the study. This consideration is similarly reported in the study from Motl et al. [75], where the correlation of average daily movement with the 6MWD is higher for a population with severe MS disability (0.47, p-value=0.001) in comparison to a population with milder symptoms (0.33, p-value=0.003). This may be due to healthy individuals engaging in a wider range of different activities in their daily routines, which makes daily movement less descriptive of physical function and capacity. Additionally, a ceiling effect may be present, where healthy subjects would all perform similarly across the identified features and tests. On the other hand, in people with mobility impairments, daily movement may more accurately reflect their actual health status and physical capabilities. This suggests that passively recorded features may be more applicable as predictors of health status in a non-healthy population, particularly where conditions affect mobility such as in neurological or cardio-pulmonary diseases.

## 5.5 Experimental protocol characteristics (RQ4)

*5.5.1 Population.* Physical function and capacity tests are usually considered within neurodegenerative diseases such as MS and PD, cardiovascular diseases, but also to monitor frailty within ageing people. This is also confirmed by the reviews from Storm et al. [100] and Pires et al. [102], where they report the majority of populations across their articles suffering from MS and cardiovascular diseases. The papers selected for review in this article (Table 2) match these populations. A third of the studies (30 %) did not consider a specific condition for participating in the data collection. This can be explained by an interest in associations across several conditions or in overall elderly frailty, such as in [69], who underwent hip replacement. It could also be the result of the early phase in which this type of research is in and, subsequently, a need for more validation across and within specific conditions. It is, for instance, feasible that different conditions have different thresholds for when an association is relevant from a clinical perspective. Additionally, the relationship between passively collected data and physical function and capacity may not provide similar results across populations where age and health vary. Only two of the selected articles addresses a paediatric population (3-12 and 5-15 years old) [89, 85], while one study includes a population of teenagers [92]. Given the different patterns that children could have compared to older adults, this is one example where additional research is needed.

*5.5.2 Protocol design.* Nearly all studies included a design with at least a visit with supervised tests and a passive monitoring period using a DHT, with certain studies following up with additional supervised visits as in [92]. This may depend highly on the specific population condition (i.e. in the work from Nair et al. this is Duchenne Muscular Dystrophy [92]). This highlights the need for a supervised standard assessment to establish the significance and relevance of data collected in free-living environments. However, standard assessments may themselves have limitations or lack comprehensive representation of participants' health, including ceiling and learning effects [103, 104]. Therefore, future studies are also needed for broad baseline assessments using multiple tests, questionnaires, and tools to establish an accurate and thorough initial understanding of each participant's condition.

The most commonly used test is the 6MWT, featured in 26 of the 47 studies. This test is widely explored and used in multiple health conditions and environments, owing to its straightforward administration and clear outcome metric (walked distance) [105]. Mueller et al. [87] note that "longer gait tests are the most reflective of real-world walking behaviour," which supports the relevance of the test for real-world application and suggests the idea of administering the test remotely in an unsupervised fashion. Several studies, such as those by Salvi et al. [106] and Ziegl et al. [107], have explored algorithms for reproducing the 6MWT in unsupervised settings. Home-based administration of the test allows for more frequent testing and in more naturalistic conditions, such as a walk in the park. The 6MWT can be augmented by processing daily life data, which can even be used to estimate an equivalent 6MWT outcome as already incorporated into the Apple Watch [55].

The 6MWT is followed by the group of X-meter walked tests (16 studies), which focus mostly on walking speed, and the TUG (12 studies). With their presence in 8 studies, the X-minute walk tests resemble the outcomes provided by the 6MWT but with different time durations. These tests are generally simple to administer for a healthcare professional, requiring minimal equipment and providing significant information related to gait, balance, and endurance of patients. The 6MWT and TUG are versatile tests used across multiple health conditions, while the Timed 25-Foot Walk test (T25FW) is particularly employed within MS [108]. While not including the activity of walking, STS-related tests are still representative of daily movements. They are reported in 7 studies and show a positive association with light intensity physical activity and moderate to vigorous physical activity [68, 64], and a negative association with sedentary behaviour [68]. This highlights the potential of this daily movement, sitting and standing, and the value of tracking it through DHT.

Not all tests reported in Table 3 are directly associated with passively collected data in the reviewed studies, but they are recognised as having the potential to be associated with free-living activities that follow similar

patterns. For example, tests such as the Hand Grip Strength test, the Arm curl or the Shuttle test were not directly associated with passively collected data in any of the included papers but appear relatively straightforward to correlate with daily physical activities that patients engage in.

*5.5.3 Monitoring duration.* Data collection in free-living conditions relies on the reliability and quality of sensor information obtained about participants' daily activities. Therefore, one necessary aspect is that participants wear the devices consistently in order to capture as much data as possible from everyday life. The duration of passive monitoring, whether shorter or longer, can significantly impact both the value of the data collected and participants' adherence to the protocol.

Within the articles selected for this review, we can see that when the monitoring period is relatively short, the study is typically aimed at capturing a snapshot of the participant's condition [58]. This may also consider the intent of investigating the feasibility of remote monitoring. In contrast, longer monitoring periods may provide more robust insights and tend to focus on tracking changes longitudinally of the participants' conditions [70, 109], and may also create a more representative insight into the physical activity patterns of participants. Challenges can arise with longer protocols for participants, however, such as reduced willingness to comply due to wearing discomfort or concerns about prolonged surveillance as longitudinal studies increasingly involve complex ethical and privacy considerations [110].

Sokas et al. [56] state that the number of detected passive 6MWT in a population of healthy control and cardiovascular disease patients increases for longer device wear time, something which became clear for a duration that exceeds 3 months. In comparison, periods that were less than a month struggled to show similar correlation levels. This deviates from the results from Zhuparris et al. [90], who conclude that 1 to 14 days are enough to estimate symptom severity in a group of people suffering from facioscapulohumeral dystrophy.

Further selected studies included in this review take a different approach by defining a minimum duration of device-wearing time within a time window (such as a single day) where data shall be considered valid. The majority of studies suggest a minimum wearing time of one week, which is similar to what is reported in the review of primarily cancer studies from Beauchamp et al. [10], who saw the most frequent duration of monitoring as between 8 to 30 days. As the examples from Sokas [56] and [90] show, optimal wearing time duration can change significantly depending on health conditions and would be a good foundation for future research to consider.

Our findings are also similar to the systematic review from Alharbi et al. [111] and to the Mobilse-D case study [112]. On one side, Alharbi et al. investigated the management of data and wearables for older adults. Similarly to our Table 6, they report that half of the articles reviewed (11/20) consider one week as the monitoring duration. On the other side, Buekers et al. had an ideal protocol of 24 hours per day, for seven days. The minimum daily wear time that did not statistically change outcomes was between no requirement to more than 14 hours of daily data, while for the number of days, it varied from 1 to more than 7 days according to different parameters to validate against clinical assessment.

*5.5.4 Protocol adherence.* Only 29/47 (62%) of the studies included in this review report on adherence. Given the relevance of wearing time in the use of passively collected data, it is of great relevance that future research provides complete information regarding compliance with their study protocol. Among those that report adherence, values are generally above 50% with the exception of [83], which used a small sample size (16 participants) and, therefore, is not comparable with the larger studies, and of [92], who reports adherence at different time intervals (1 week, 1 year and 2 year), clearly observing the problems arising with longer monitoring or follow-up periods. For a one-week monitoring period (20 studies), the adherence is above 57%, with an average of 82%, which gives further support for week-long monitoring as a good minimum target. Two studies achieve adherence of 98% [75] and 94% [69] even though they include a large group of participants (256 and 239, respectively). However, their minimum

monitoring time is very permissive as they only require a minimum of three days of data collection from each participant to be counted as compliant.

Two studies that use one year of monitoring obtained 57% [59] and 83% [73] adherence. In longitudinal monitoring, one may expect low adherence to the study protocol but other factors also affect this, such as the type of recruitment (common dweller participants not followed by clinical experts [59], or recruited MS patients through a clinical facility [73]), and the type and position of devices used (unobtrusive passive ambient monitoring devices [59], or using a device with highly validated form factor [73]). During real-world data collection over a longer period of time, protocol adherence is reported as improved by contacting participants through phone calls, as for example, in [70]. Mueller et al. [87] illustrate participants' adherence, highlighting the weak compliance of participants throughout their study time and the complexity of managing adherence.

A systematic review of 25 articles on the use of wearables for cancer treatment reports an adherence for patients wearing time between 60% to 100%, and 45%-94% of valid days [10]. The adherence values reported in this review are relatively higher. This could be given the fact that the sample sizes of these articles are smaller in comparison to the ones in this work (7-180 participants vs 10-703) and the fact that cancer patients may be more engaged in their treatment. Their review, in agreement with our findings, reports that adherence measurement is inconsistent across studies and, thus, challenging to compare. In the study from Sokas et al. [56], researchers asked participants to wear a FitBit device for at least one week. Reported adherence is 71.08% in terms of the days ratio. In addition, 24/28 cardiovascular disease patients and 67/71 healthy subjects exceeded the monitoring duration of one week, while 16/28 patients and 62/71 healthy subjects exceeded two weeks. Some participants in this study even exceeded the monitoring duration of 1 and 3 months. This is promising and could highlight the fact that a good portion of participants are generally willing to use wearable devices.

## 6 Conclusions

In this review, we aim to provide a comprehensive summary of current research on the use of sensors in free-living conditions and their clinical significance when associated with standard physical tests. To do so, we defined four research questions that guided our systematic review of 47 existing studies, in which we outline key results below - including considerations for future research related to the respective research questions. The studies were identified following the PRISMA [48] search guidelines, targeting the Scopus, PubMed, and Web of Science databases in combination with snowballing based on references analysis.

**RQ1: Associating passively collected data with physical tests.** Yes, it is possible to associate data collected by DHT in free-living environments with standard physical tests, as many studies demonstrate significant correlations between these data types. However, exceptions exist where no significant relationship between free-living physical activity and clinical tests was observed. Also, while correlations and classifications provide clinical significance to passively collected data, further research is needed to confirm these associations across different populations and health conditions and to validate their use in clinical practice.

**RQ2: Digital health technologies.** The most common DHT used for passive data collection are wrist-worn devices with IMUs from brands like Fitbit, Apple Watch, Withings, and ActiGraph, which is also in line with the results from other reviews. As the included studies do not compare the data quality with other devices, it remains unclear if the specific brand or model is more or less relevant than other commonly available wearables. Smartphones also show promise for data collection due to their accessibility and widespread use across demographics. Other sensing modalities, such as localization, have the potential for increasing accuracy and insights, but their impact on privacy and ethical aspects would also need to be considered.

**RQ3: Metrics extraction and their association with physical tests.** Metrics extracted from daily living data are most commonly derived from step count and activity intensity levels. Temporal features such as time spent in moderate-to-vigorous activity are also prominent, with measures like the NET-F index and Peak-30CAD

showing correlations up to 0.68 [94, 76]. Overall, correlations between passive data and clinical tests tend to be stronger in populations with physical impairments, suggesting that such features may more accurately reflect the physical function of individuals with health conditions rather than healthy participants.

**RQ4: Experimental protocol characteristics.** The reviewed studies share several common characteristics in their experimental protocols, particularly regarding population, physical tests, monitoring duration, and protocol adherence.

- Population: Studies primarily focus on populations affected by neurodegenerative diseases (e.g., multiple sclerosis, Parkinson's disease), cardiovascular conditions, or frailty due to ageing, as these conditions impact physical function and capacity. However, numerous studies did not specify a particular health condition, likely aiming to explore passively collected data in a general population. The risk of bias analysis highlighted challenges related to the recruitment of representative groups of participants, with the tendency to recruit people with a higher level of fitness in comparison to a baseline target of interest. Only three studies included a population below 15 years old, indicating the need for further exploration in younger age groups.
- Protocol Design: Most studies are characterised by cross-sectional approaches, including a supervised assessment visit and a passive monitoring period with DHT. Supervised assessments serve as baseline measures, helping validate data collected in unsupervised settings. To note is that cross-sectional designs allow establishing statistical correlation rather than causality between performance and capacity. The 6MWT is the most frequently used test, appearing in 26 studies. Future studies would benefit from using multiple tests and questionnaires to establish a thorough baseline of participants' conditions.
- Monitoring Duration: Most studies suggest a minimum wearing time of one week, aligning with other reviews on wearables in health research. Optimal wearing time, however, may vary across conditions, as seen in studies specifying different minimum durations based on health contexts.
- Protocol Adherence: Short-term monitoring achieves generally higher adherence, with participants more likely to maintain the protocol over days or weeks. Long-term studies report lower adherence, which may be mitigated by contacting participants over the phone but remains an issue which needs to be actively addressed. In some of the selected studies, participants exceeded the required wear time, suggesting that user engagement with wearable devices can be very high if the DHT provides tangible user value.

Our review has some limitations, which are useful to consider for future research.

First, even if a broad query on physical function and capacity was included, our focus was on 4 types of standardised physical tests. Regardless of how representative and widely adopted those tests are, there are more physical tests that can be included, such as the T25FW, which we did not include in the query but appeared as relevant for MS in the snowballing phase.

Second, the clinical relevance of data collected in free-living conditions can be assessed in other ways rather than with physical tests, such as by assessing its association with mortality or by measuring response to medication and treatments in general. These studies, however, require large cohorts and time spans, which may not be easy to obtain with such recent technologies.

Third, the reviewed papers focussed mostly on non-communicable diseases, particularly chronic conditions that are known to affect mobility such as neurodegenerative disease, however there may be valuable information in free living activity data also for other types of conditions, including infectious diseases. For example, a reduction in physical activity measured through wearables has been associated with the onset of COVID-19 [113] or influenza [114]. This could be the subject of future research.

Finally, future work could also examine how categories of physical activity relate to standard test performance (high vs low performance), emphasizing their use in clinical practice and decision-making. This information may

usefully complement other assessments and motivate further research on finer classifications using time series analysis with machine learning and advanced techniques.

Overall, our review contributes by adding to the evidence that sensors and digital technologies are becoming valuable tools for assessing human health. In the future, these tools may become more effective and timely at detecting changes than conventional clinical tests and may become a reliable tool for aiding diagnosis and treatments.

## Acknowledgments

This research is partially funded by the Swedish Knowledge Foundation and the Sustainable Digitalization Research Center (F.k.a. Internet of Things and People Research Center) through the Intelligent and Trustworthy IoT Systems project, and it is also co-funded by the European Union - Next Generation EU, under the National Recovery and Resilience Plan, Investment Partenariato Esteso PE8 'Conseguenze e sfide dell'invecchiamento,' Project Age-IT, CUP: B83C22004800006.

## References

[1] Lukasz Piwek et al. "The rise of consumer health wearables: promises and barriers". In: *PLoS medicine* 13.2 (2016), e1001953.

[2] Luisa Barrera-Leon et al. "Development of a Support System for Physicians and Patients during Rehabilitation". In: *Biomechanics* 4.3 (2024), pp. 520–541.

[3] Diane Podsiadlo and Sandra Richardson. "The timed "Up & Go": a test of basic functional mobility for frail elderly persons". In: *Journal of the American geriatrics Society* 39.2 (1991), pp. 142–148.

[4] Sarah E Lamb and David J Keene. "Measuring physical capacity and performance in older people". In: *Best practice & research Clinical rheumatology* 31.2 (2017), pp. 243–254.

[5] Lawrence P Cahalin et al. "The six-minute walk test predicts peak oxygen uptake and survival in patients with advanced heart failure". In: *Chest* 110.2 (1996), pp. 325–332.

[6] Ya-Ting Liang, Charlotte Wang, and Chuhsing Kate Hsiao. "Data Analytics in Physical Activity Studies With Accelerometers: Scoping Review". In: *Journal of Medical Internet Research* 26 (2024), e59497.

[7] Kanika Bansal et al. "Speed-and Endurance-Based Classifications of Community Ambulation Post-Stroke Revisited: The Importance of Location in Walking Performance Measurement". In: *Neurorehabilitation and Neural Repair* (2024), p. 15459683241257521.

[8] Jenna A Zajac et al. "Does clinically measured walking capacity contribute to real-world walking performance in Parkinson's disease?" In: *Parkinsonism & related disorders* 105 (2022), pp. 123–127.

[9] Jason Weatherald et al. "Clinical trial design, end-points, and emerging therapies in pulmonary arterial hypertension". In: *European Respiratory Journal* (2024).

[10] Ulrikke Lyng Beauchamp, Helle Pappot, and Cecilie Holländer-Mieritz. "The use of wearables in clinical trials during cancer treatment: systematic review". In: *JMIR mHealth and uHealth* 8.11 (2020), e22006.

[11] Graeme Mattison et al. "The influence of wearables on health care outcomes in chronic disease: systematic review". In: *Journal of Medical Internet Research* 24.7 (2022), e36690.

[12] Jose Oscar Olmedo-Aguirre et al. "Remote healthcare for elderly people using wearables: A review". In: *Biosensors* 12.2 (2022), p. 73.

[13] Marie Chan et al. "Smart wearable systems: Current status and future challenges". In: *Artificial intelligence in medicine* 56.3 (2012), pp. 137–156.

[14] Mohammad Moshawrab et al. "Smart wearables for the detection of cardiovascular diseases: a systematic literature review". In: *Sensors* 23.2 (2023), p. 828.

[15] Marco Giurgiu et al. "Quality evaluation of free-living validation studies for the assessment of 24-hour physical behavior in adults via wearables: systematic review". In: *JMIR mHealth and uHealth* 10.6 (2022), e36377.
[16] Mariano Bernaldo de Quiros et al. "Quantification of Movement in Stroke Patients under Free Living Conditions Using Wearable Sensors: A Systematic Review". In: *Sensors* 22.3 (2022), p. 1050.
[17] Sheikh MA Iqbal et al. "Advances in healthcare wearable devices". In: *NPJ Flexible Electronics* 5.1 (2021), p. 9.
[18] Ali K Yetisen et al. "Wearables in medicine". In: *Advanced Materials* 30.33 (2018), p. 1706910.
[19] Rodica Lucia Avram et al. "Functional tests in patients with ischemic heart disease". In: *Journal of Medicine and Life* 15.1 (2022), p. 58.
[20] Enrica Patrizio et al. "Physical functional assessment in older adults". In: *The Journal of frailty & aging* 10 (2021), pp. 141–149.
[21] Carl J Caspersen, Kenneth E Powell, and Gregory M Christenson. "Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research." In: *Public health reports* 100.2 (1985), p. 126.
[22] Daniel E Forman et al. "Prioritizing functional capacity as a principal end point for therapies oriented to older adults with cardiovascular disease: a scientific statement for healthcare professionals from the American Heart Association". In: *Circulation* 135.16 (2017), e894–e918.
[23] Jacqueline M Miotto et al. "Reliability and validity of the Fullerton Functional Fitness Test: an independent replication study". In: *Journal of Aging and Physical activity* 7.4 (1999), pp. 339–353.
[24] Kenneth J Ottenbacher et al. "The reliability of the functional independence measure: a quantitative review". In: *Archives of physical medicine and rehabilitation* 77.12 (1996), pp. 1226–1232.
[25] Luc JW Evers et al. "Measuring Parkinson's disease over time: the real-world within-subject reliability of the MDS-UPDRS". In: *Movement Disorders* 34.10 (2019), pp. 1480–1487.
[26] Cristiana Lopes Gabriel et al. "Mobile and wearable technologies for the analysis of Ten Meter Walk Test: A concise systematic review". In: *Heliyon* 9.6 (2023).
[27] Kelley K Pettee Gabriel et al. "Test-retest reliability and validity of the 400-meter walk test in healthy, middle-aged women". In: *Journal of Physical Activity and Health* 7.5 (2010), pp. 649–657.
[28] Anne B Newman et al. "Association of long-distance corridor walk performance with mortality, cardiovascular disease, mobility limitation, and disability". In: *Jama* 295.17 (2006), pp. 2018–2026.
[29] Rengin Demir and Mehmet Serdar Küçükoğlu. "Six-minute walk test in pulmonary arterial hypertension". In: *Anatolian journal of cardiology* 15.3 (2015), p. 249.
[30] Ivan Bautmans, Margareta Lambert, and Tony Mets. "The six-minute walk test in community dwelling elderly: influence of health status." In: *BMC geriatrics* 4 (2004), pp. 1–9.
[31] Sema Savci et al. "Six-minute walk distance as a measure of functional exercise capacity in multiple sclerosis". In: *Disability and rehabilitation* 27.22 (2005), pp. 1365–1371.
[32] Andrey Jorge Serra et al. "Correlation of six-minute walking performance with quality of life is domain-and gender-specific in healthy older adults". In: *PLoS one* 10.2 (2015), e0117359.
[33] Priya Agarwala and Steve H Salzman. "Six-minute walk test: clinical role, technique, coding, and reimbursement". In: *Chest* 157.3 (2020), pp. 603–611.
[34] Susan Morris, Meg E Morris, and Robert Iansek. "Reliability of measurements obtained with the Timed "Up & Go" test in people with Parkinson disease". In: *Physical therapy* 81.2 (2001), pp. 810–818.
[35] Paulina Ortega-Bastidas et al. "Instrumented timed up and go test (itug)—More than assessing time to predict falls: A systematic review". In: *Sensors* 23.7 (2023), p. 3426.
[36] Pei Ling Choo et al. "Timed Up and Go (TUG) reference values and predictive cutoffs for fall risk and disability in Singaporean community-dwelling adults: Yishun cross-sectional study and Singapore

longitudinal aging study". In: *Journal of the American Medical Directors Association* 22.8 (2021), pp. 1640–1645.

[37] Teresa M Steffen, Timothy A Hacker, and Louise Mollinger. "Age-and gender-related test performance in community-dwelling elderly people: Six-Minute Walk Test, Berg Balance Scale, Timed Up & Go Test, and gait speeds". In: *Physical therapy* 82.2 (2002), pp. 128–137.

[38] Trija Vaidya, Arnaud Chambellan, and Claire De Bisschop. "Sit-to-stand tests for COPD: a literature review". In: *Respiratory medicine* 128 (2017), pp. 70–77.

[39] Ryan P Duncan, Abigail L Leddy, and Gammon M Earhart. "Five times sit-to-stand test performance in Parkinson's disease". In: *Archives of physical medicine and rehabilitation* 92.9 (2011), pp. 1431–1436.

[40] Richard W Bohannon. "Sit-to-stand test for measuring performance of lower extremity muscles". In: *Perceptual and motor skills* 80.1 (1995), pp. 163–166.

[41] Kevin Sykes and Alison Roberts. "The Chester step test—a simple yet effective tool for the prediction of aerobic capacity". In: *Physiotherapy* 90.4 (2004), pp. 183–188.

[42] Anderson Alves de Camargo et al. "Chester step test in patients with COPD: reliability and correlation with pulmonary function test results". In: *Respiratory care* 56.7 (2011), pp. 995–1001.

[43] World Health Organization et al. "IFC: International classification of functioning, disability and health". In: (2001).

[44] Catherine E Lang et al. "Improvement in the capacity for activity versus improvement in performance of activity in daily life during outpatient rehabilitation". In: *Journal of Neurologic Physical Therapy* 47.1 (2023), pp. 16–25.

[45] Nuttawuth Mekritthikrai, Kornanong Yuenyongchaiwat, and Chusak Thanawattano. "Concurrent validity and reliability of new application for 6-min walk test in healthy adults". In: *Heliyon* 9.7 (2023).

[46] Jonathan Mak et al. "Reliability and repeatability of a smartphone-based 6-min walk test as a patient-centred outcome measure". In: *European Heart Journal-Digital Health* 2.1 (2021), pp. 77–87.

[47] Cormac Gerard Ryan et al. "The convergent validity of free-living physical activity monitoring as an outcome measure of functional ability in people with chronic low back pain". In: *Journal of Back and Musculoskeletal Rehabilitation* 21.2 (2008), pp. 137–142.

[48] Matthew J Page et al. "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews". In: *bmj* 372 (2021).

[49] Jonathan AC Sterne et al. "Assessing risk of bias in a non-randomized study". In: *Cochrane handbook for systematic reviews of interventions* (2019), pp. 621–641.

[50] *Anara*. https://anara.com/. Accessed: December 2025.

[51] Joana Silva, Diana Gomes, and Inês Sousa. "Technological solution for pervasive fall risk assessment". In: *pHealth 2020*. IOS Press, 2020, pp. 176–181.

[52] Marilyn L Moy et al. "Free-living physical activity in COPD: assessment with accelerometer and activity checklist". In: *Journal of rehabilitation research and development* 46.2 (2009), p. 277.

[53] Nicola Straiton et al. "Wearable activity trackers objectively measure incidental physical activity in older adults undergoing aortic valve replacement". In: *Sensors* 23.6 (2023), p. 3347.

[54] Irene Katzan, Andrew Schuster, Tyler Kinzy, et al. "Physical activity monitoring using a fitbit device in ischemic stroke patients: prospective cohort feasibility study". In: *JMIR mHealth and uHealth* 9.1 (2021), e14494.

[55] *Using Apple Watch to Estimate Six-Minute Walk Distance*. https://www.apple.com/healthcare/docs/site/Using_Apple_Watch_to_Estimate_Six_Minute_Walk_Distance.pdf. Accessed: November 2024. 2021.

[56] Daivaras Sokas et al. "Detection of walk tests in free-living activities using a wrist-worn device". In: *Frontiers in physiology* 12 (2021), p. 706545.

[57] Breanna M Crane et al. "Using GPS technologies to examine community mobility in older adults". In: *The Journals of Gerontology: Series A* 78.5 (2023), pp. 811–820.
[58] Julie Rekant et al. "Physical Functioning, Physical Activity, and Variability in Gait Performance during the Six-Minute Walk Test". In: *Sensors* 24.14 (2024), p. 4656.
[59] Narayan Schütz et al. "A systems approach towards remote health-monitoring in older adults: Introducing a zero-interaction digital exhaust". In: *NPJ digital medicine* 5.1 (2022), p. 116.
[60] Antti Löppönen et al. "Association of sit-to-stand capacity and free-living performance using Thigh-Worn accelerometers among 60-to 90-yr-old adults". In: *Medicine and Science in Sports and Exercise* 55.9 (2023), p. 1525.
[61] Emily Wright, Victoria Chester, and Usha Kuruganti. "Identifying the Optimal Parameters to Express the Capacity–Activity Interrelationship of Community-Dwelling Older Adults Using Wearable Sensors". In: *Sensors* 22.24 (2022), p. 9648.
[62] Diana A Santos et al. "Sedentary behavior and physical activity are independently related to functional fitness in older adults". In: *Experimental gerontology* 47.12 (2012), pp. 908–912.
[63] Jacek K Urbanek et al. "Validation of gait characteristics extracted from raw accelerometry during walking against measures of physical function, mobility, fatigability, and fitness". In: *The Journals of Gerontology: Series A* 73.5 (2018), pp. 676–681.
[64] MYLES W O'BRIEN et al. "Validation of the PiezoRx® step count and moderate to vigorous physical activity times in free living conditions in adults: a pilot study". In: *International journal of exercise science* 11.7 (2018), p. 541.
[65] Maaike van Gameren et al. "The Short Physical Performance Battery does not correlate with daily life gait quality and quantity in community-dwelling older adults with an increased fall risk". In: *Gait & posture* 114 (2024), pp. 78–83.
[66] Laura Karavirta et al. "Individual scaling of accelerometry to preferred walking speed in the assessment of physical activity in older adults". In: *The Journals of Gerontology: Series A* 75.9 (2020), e111–e118.
[67] Tatiane Lopes de Pontes et al. "Total energy expenditure and functional status in older adults: a doubly labelled water study". In: *The Journal of nutrition, health and aging* 25.2 (2021), pp. 201–208.
[68] Renoa Choudhury et al. "Objectively measured physical activity levels and associated factors in older US women during the COVID-19 pandemic: cross-sectional study". In: *JMIR aging* 5.3 (2022), e38172.
[69] Salvatore Saporito et al. "Remote timed up and go evaluation from activities of daily living reveals changing mobility after surgery". In: *Physiological Measurement* 40.3 (2019), p. 035004.
[70] Shaoxiong Sun et al. "The utility of wearable devices in assessing ambulatory impairments of people with multiple sclerosis in free-living conditions". In: *Computer methods and programs in biomedicine* 227 (2022), p. 107204.
[71] Yuyang Zhai et al. "Smartphone accelerometry: A smart and reliable measurement of real-life physical activity in multiple sclerosis and healthy individuals". In: *Frontiers in neurology* 11 (2020), p. 688.
[72] VJ Block et al. "Continuous daily assessment of multiple sclerosis disability using remote step count monitoring". In: *Journal of neurology* 264 (2017), pp. 316–326.
[73] Valerie J Block et al. "Association of continuous assessment of step count by remote monitoring with disability progression among adults with multiple sclerosis". In: *JAMA network open* 2.3 (2019), e190570–e190570.
[74] Akara Supratak et al. "Remote monitoring in the home validates clinical gait measures for multiple sclerosis". In: *Frontiers in neurology* 9 (2018), p. 561.
[75] RW Motl et al. "Accelerometry as a measure of walking behavior in multiple sclerosis". In: *Acta Neurologica Scandinavica* 127.6 (2013), pp. 384–390.

[76] Peixuan Zheng et al. "Free-living peak cadence in multiple sclerosis: a new measure of real-world walking?" In: *Neurorehabilitation and Neural Repair* 37.10 (2023), pp. 716–726.
[77] Kedar KV Mate and Nancy E Mayo. "Clinically assessed walking capacity versus real-world walking performance in people with multiple sclerosis". In: *International Journal of MS Care* 22.3 (2020), pp. 143–150.
[78] Viktoria Karle et al. "The two-minute walk test in persons with multiple sclerosis: correlations of cadence with free-living walking do not support ecological validity". In: *International Journal of Environmental Research and Public Health* 17.23 (2020), p. 9044.
[79] Neil Rens et al. "Activity data from wearables as an indicator of functional capacity in patients with cardiovascular disease". In: *Plos one* 16.3 (2021), e0247834.
[80] Ségolène Chaudru et al. "Using wearable monitors to assess daily walking limitations induced by ischemic pain in peripheral artery disease". In: *Scandinavian Journal of Medicine & Science in Sports* 29.11 (2019), pp. 1813–1826.
[81] Alex V Rowlands et al. "Can quantifying the relative intensity of a person's free-living physical activity predict how they respond to a physical activity intervention? Findings from the PACES RCT". In: *British Journal of Sports Medicine* 57.22 (2023), pp. 1428–1434.
[82] Martin Ullrich et al. "Detection of unsupervised standardized gait tests from real-world inertial sensor data in Parkinson's disease". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 29 (2021), pp. 2103–2111.
[83] Raquel Bouça-Machado et al. "Kinematic and clinical outcomes to evaluate the efficacy of a multidisciplinary intervention on functional mobility in Parkinson's disease". In: *Frontiers in neurology* 12 (2021), p. 637620.
[84] Håkan Nero et al. "Objectively assessed physical activity and its association with balance, physical function and dyskinesia in Parkinson's disease". In: *Journal of Parkinson's Disease* 6.4 (2016), pp. 833–840.
[85] M Nørgaard et al. "Accelerometry-based monitoring of daily physical activity in children with juvenile idiopathic arthritis". In: *Scandinavian Journal of Rheumatology* 45.3 (2016), pp. 179–187.
[86] Vibhu Agarwal et al. "Inferring physical function from wearable activity monitors: analysis of free-living activity data from patients with knee osteoarthritis". In: *JMIR mHealth and uHealth* 6.12 (2018), e11315.
[87] Arne Mueller et al. "Continuous digital monitoring of walking speed in frail elderly patients: noninterventional validation study and longitudinal clinical trial". In: *JMIR mHealth and uHealth* 7.11 (2019), e15191.
[88] Ingrid van de Port, Michiel Punt, and Jan Willem Meijer. "Walking activity and its determinants in free-living ambulatory people in a chronic phase after stroke: a cross-sectional study". In: *Disability and rehabilitation* 42.5 (2020), pp. 636–641.
[89] Isabella Wiedmann et al. "Accelerometric gait analysis devices in children—will they accept them? results from the AVAPed study". In: *Frontiers in Pediatrics* 8 (2021), p. 574443.
[90] Ahnjili Zhuparris et al. "Smartphone and wearable sensors for the estimation of facioscapulohumeral muscular dystrophy disease severity: cross-sectional study". In: *JMIR Formative Research* 7.1 (2023), e41178.
[91] S Birnbaum et al. "Free-living physical activity and sedentary behaviour in auto-immune myasthenia gravis: a cross-sectional study". In: *Neuromuscular Disorders*. Vol. 30. PERGAMON-ELSEVIER SCIENCE LTD THE BOULEVARD, LANGFORD LANE, KIDLINGTON … 2020, S59–S59.
[92] Kavya S Nair et al. "Step activity monitoring in boys with Duchenne muscular dystrophy and its correlation with magnetic resonance measures and functional performance". In: *Journal of neuromuscular diseases* 9.3 (2022), pp. 423–436.
[93] Kate Lyden et al. "Predicting hospitalization from real-world measures in patients with chronic kidney disease: A proof-of-principle study". In: *Digital Health* 9 (2023), p. 20552076231181234.

[94] Alessandra Angelucci et al. "Fitbit data to assess functional capacity in patients before elective surgery: pilot prospective observational study". In: *Journal of Medical Internet Research* 25.1 (2023), e42815.
[95] Carissa A Low et al. "Associations between performance-based and patient-reported physical functioning and real-world mobile sensor metrics in older cancer survivors: a pilot study". In: *Journal of Geriatric Oncology* 15.2 (2024), p. 101708.
[96] Leena Choi et al. "Validation of accelerometer wear and nonwear time classification algorithm". In: *Medicine and science in sports and exercise* 43.2 (2011), p. 357.
[97] Shelby L Bachman et al. "Do Measures of Real-World Physical Behavior Provide Insights Into the Well-Being and Physical Function of Cancer Survivors? Cross-Sectional Analysis". In: *JMIR cancer* 10 (2024), e53180.
[98] *Wearable Technology Market Analysis By Product, By Application And Segment Forecasts From 2015 To 2022*. https://www.millioninsights.com/industry-reports/wearable-technology-market. Report number: MN17617489, Accessed: November 2024. 2022.
[99] *Europe Digital Health Market Size, Share & Trends Analysis Report By Technology (Tele-healthcare, mHealth), By Component (Software, Hardware), By Application, By End-use, By Country, And Segment Forecasts, 2024 - 2030*. https://www.grandviewresearch.com/industry-analysis/europe-digital-health-market-report. Report number: GVR-4-68039-941-1, Accessed: November 2024.
[100] Fabio Alexander Storm et al. "Wearable inertial sensors to assess gait during the 6-minute walk test: A systematic review". In: *Sensors* 20.9 (2020), p. 2660.
[101] Martin Kögler et al. "Sustainable use of a smartphone and regulatory needs". In: *Sustainable Development* (2024).
[102] Ivan Miguel Pires et al. "Development technologies for the monitoring of six-minute walk test: a systematic review". In: *Sensors* 22.2 (2022), p. 581.
[103] Grace Wu, Bonnie Sanderson, and Vera Bittner. "The 6-minute walk test: how important is the learning effect?" In: *American heart journal* 146.1 (2003), pp. 129–133.
[104] Talia Herman, Nir Giladi, and Jeffrey M Hausdorff. "Properties of the 'timed up and go'test: more than meets the eye". In: *Gerontology* 57.3 (2011), pp. 203–210.
[105] Richard W Bohannon and Rebecca Crouch. "Minimal clinically important difference for change in 6-minute walk test distance of adults with pathology: a systematic review". In: *Journal of evaluation in clinical practice* 23.2 (2017), pp. 377–381.
[106] Dario Salvi et al. "The mobile-based 6-minute walk test: usability study and algorithm development and validation". In: *JMIR mHealth and uHealth* 8.1 (2020), e13756.
[107] Andreas Ziegl et al. "mHealth 6-minute walk test–accuracy for detecting clinically relevant differences in heart failure patients". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 7095–7098.
[108] Robert W Motl et al. "Validity of the timed 25-foot walk as an ambulatory performance outcome measure for multiple sclerosis". In: *Multiple Sclerosis Journal* 23.5 (2017), pp. 704–710.
[109] Andong Zhan et al. "Using smartphones and machine learning to quantify Parkinson disease severity: the mobile Parkinson disease score". In: *JAMA neurology* 75.7 (2018), pp. 876–880.
[110] Anna Sui et al. "Ethical considerations for the use of consumer wearables in health research". In: *Digital Health* 9 (2023), p. 20552076231153740.
[111] Muaddi Alharbi et al. "Data management and wearables in older adults: A systematic review". In: *Maturitas* 124 (2019), pp. 100–110.
[112] Joren Buekers et al. "Digital assessment of real-world walking in people with impaired mobility: How many hours and days are needed?" In: *International Journal of Behavioral Nutrition and Physical Activity* 22.1 (2025), p. 148.

[113] Asma Channa et al. “The rise of wearable devices during the COVID-19 pandemic: A systematic review”. In: *Sensors* 21.17 (2021), p. 5787.
[114] Kamran Farooq et al. “Evaluation of Machine Learning to Detect Influenza Using Wearable Sensor Data and Patient-Reported Symptoms: Cohort Study”. In: *Journal of Medical Internet Research* 26 (2024), e47879.

## A Appendix

Table 7 reports outcomes from each paper reviewed, while Table 8 shows the risk of bias assessment.

Table 7. Relationship between data collected in free-living conditions and standard clinical assessment.

| Article | Found association | Device Position | Type of sensor | Monitoring duration | Considered tests | Participants adherence |
|---|---|---|---|---|---|---|
| Wiedmann et al. [89] | Gait speed and 1MWT (0.26, $p = 0.002$). Gait speed and 1MWT gait speed (0.40, $p = 0.004$) | Waist | IMU | 1 week | 1MWT | 69.32% |
| Rens et al. [79] | Passive data ability to distinguish an in-clinic 6MWT distance higher or lower 300m with AUC 0.643, and with unsupervised reproduced 6MWT distance higher or lower 300m with AUC 0.704 (natural logarithm of the total steps walked per day as counted by the phone or watch, whichever was higher) | Undefined + Wrist | IMU, PPG | 6 months | 6MWT | 72.90% |
| Mueller et al. [87] | Gait speed in short bouts and 4mWT gait speed (0.23), longer bouts correlated with 6MWT (0.48) and with 400mWT (0.59). Weaker associations for comparison of non-similar walking bouts. Longer gait tests are the most reflective of real-world walking behaviour. | Waist | IMU | 40 days | 4MWT, 6MWT, 400mWT | |
| Ullrich et al. [82] | Good agreement between gait parameters derived from detected 4 × 10MWT in free-living conditions compared to ground truth 4 × 10MWT labels was observed for all gait parameters in all speed levels. | Shoes | IMU | 2 weeks | 10mWT | |
| Sokas et al. [56] | Correlation between estimated 6MWD and reference 6MWD for patients with CVD, healthy subjects over 40 years, and healthy subjects below 40 years, is 0.39 ($p < 0.01$), 0.23 (p=0.07), and 0.42 ($p < 0.05$), respectively. Unintentional walk testing is feasible and could be valuable for repeated assessment of functional performance outside the clinical setting. | Wrist | IMU, PPG | 1 week | 6MWT | 71,08% |
| Lyden et al. [93] | Best 6-minute step count correlated with in-clinic assessments as follows: with 6MWD (0.26, p=0.007), with 4mWT (0.20, p=0.04), with shuttle walk test (0.29, p=0.003). Rise time correlated with in-clinic assessments as follows: with 6MWD (-0.26, p=0.01), with 4mWT (-0.23, p=0.02), with shuttle walk test (-0.22, p=0.03). Sedentary duration was not correlated with in-clinic physical function measurements, while MVPA and the number of steps/day had modest correlation with in-clinic 6-minute walk distance. | Wrist | IMU, PPG | 1 week | 6MWT, Shuttle Test, 4mWT | |
| Sun et al. [70] | Minute-level step count had a stronger association with 6MWT. Best features are walking aids, age, sedentary activity, max steps in short bouts (up to 30 min). Classification between upper and lower 25% 6MWD derived from maximum, AUC of random forest = 0.87 | Undefined + Wrist | IMU, PPG, Smartphone interaction | 6 to 24 months | 6MWT | |
| Crane et al. [57] | Standard deviational ellipse (SDE) area and 6MWT (0.24, $p = 0.003$), Total time and percentage of total time out-of-home and 6MWT (0.22-0.27, $p = .01$). Total time and percentage of total time out-of-home and time to complete Figure 8 Walk (-0.20–0.23, $p = .05$). Median time out-of-home was correlated with 6MWT ($\rho = 0.24$, $p = .003$). The overall maximum distance traveled from home was associated with better performance on the Trail Making Test, Part B ($\rho = -0.21$, $p = .01$), while the median maximum distance was related to faster times on the Trail Making Test, Part B ($\rho = -0.18$, $p = .03$), a greater distance traveled on the 6MWT ($\rho = 0.24$, $p = .004$), and faster times on the Figure of 8 Walk ($\rho = -0.16$, $p = .047$). | Undefined | GPS | 1 week | 6MWT, Figure 8 Walk Test, 14mWT | 81,87% |
| Straiton et al. [53] | Improved functional capacity post-AVR in the clinic setting does not always correspond to improved engagement in incidental PA in the free-living environment and vice versa. | Wrist | IMU | 8 days | 6MWT, 5mWT, Hand Grip Strength Test | 73.33% |
| Rekant et al. [58] | Neither summary measures of global health nor physical activity were significantly related to 6MWT performance. | Wrist | IMU | 1 day | FSST, 10mWT, 6MWT | |
| Zajac et al. [8] | Clinically measured walking capacity significantly contributed to real-world walking performance (i.e., daily steps and weekly moderate intensity walking minutes) in a sample of relatively older persons with mild to moderate PD. However, a large portion of the variance in walking performance was unexplained by capacity measure. Nonetheless, walking capacity explained more variance in a less active subgroup compared to a more active subgroup of participants. | Ankle | IMU, GPS | 1 week | 6MWT, 10mWT | |
| Schütz et al. [59] | Digital exhaust performances on the fall-risk related TUG has ROC AUC values of 0.786 | Undefined | Ambient sensors | 1 year | TUG | 57,14% |

*Continued on next page*

| Article | Found association | Device Position | Type of sensor | Monitoring duration | Considered tests | Participants adherence |
|---|---|---|---|---|---|---|
| Bouça-Machado et al. [83] | The analysis using supervised and FL kinematic gait parameters as independent variables identified step length (adjusted R2 = 0.53) and step time asymmetry (adjusted R2 = 0.51) as the best predictors of TUG changes for each assessment condition. | Lower back | IMU | 6 days | TUG, 5STS | 66,67% |
| Silva et al. [51] | Both theoretical (parameters weights based on literature odds ratio) and empirical (parameters weights based on correlation of each one with TUG and POMA tests) approaches presented strong correlations with the functional tests (TPOMA, 0.80; T-TUG, 0.79; E-POMA, 0.90; E-TUG, 0.89), and were deemed statistically similar to the scaled output of these assessments, presenting p-values $< 0.1$. | Undefined | IMU | 2 weeks | TUG, POMA | |
| Zhai et al. [71] | Steps/minute showed within all participants weak to moderate correlations with clinical measures (2MWT, 6MWT, FTSTS,T25FW rho = $\|0.21\|to\|0.34\|$, $p < 0.05$). In healthy controls and pwMS with ambulatory impairment, MVPA had a stronger association with some of the clinical measures than steps/minute (rho = $\|0.28\|to\|0.59\|$, $p < 0.05$). Among all variables derived from the smartphone, varVM showed the strongest correlations among all participants with the clinical measures (TTW, 2MWT, 6MWT, FTSTS, T25FW, rho = $\|0.56\|to\|0.61\|$, $p < 0.0001$). | Undefined + Wrist | IMU | 1 week | 5STS, T25FW, 2MWT, 6MWT, 3mTTW | 86,71% |
| Block et al. [72] | Average daily step count was associated with T25FW times (Spearman's q = -0.65, $p < 0.001$). Similarly, TUG times and distances walked during the 2MW test were also significantly associated with step count. | Wrist | IMU | 1 week | T25FW, TUG, 2MWT | |
| Zhuparris et al. [90] | Time spent at a health-related location (such as a gym or hospital) and call duration were features that were predictive of clinical assessments. The multitask model had R2 of 0.81 and RMSE of 1.61 for the TUG. The 3 most important features were light sleep duration, total steps per day, and mean steps per minute. Using an increasing time window (starting from day 1 to day 14) for the TUG estimation yielded an average R2 of 0.79 and an average RMSE of 2.05. | Wrist | IMU, PPG | 6 weeks | TUG | 65,52% |
| Block et al. [73] | A decreasing average daily step count during the study was associated with worsening of clinic-based outcomes (Timed 25-Foot Walk, $\beta = -13.09$; $P < .001$; Timed-Up-and-Go, $\beta = -9.25$; $P < .001$). | Wrist | IMU | 1 year | T25FW, 6MWT, 10MWT, 2MWT | 83,16% |
| Supratak et al. [74] | Correlation between maximum sustained walking speeds at home and the T25FW walking speed measured in the clinic R-value = 0.89. | Lower back | IMU | 1 week | T25FW, 6MWT | |
| Motl et al. [75] | The average of total daily movement counts from the vertical axis of the accelerometer correlated with T25FW ($\rho = -0.595$, $P = 0.001$) and 6MWD ($\rho = 0.630$, $P = 0.001$). Among those with mild disability, the correlations resulted in $\rho = -0.386$, $P = 0.001$ and $\rho = 0.333$, $P = 0.003$. Among those with moderate/severe disability, the correlations resulted in ($\rho = -0.431$, $P = 0.001$) and ($\rho = 0.469$, $P = 0.001$). | Undefined | IMU | 1 week | T25FW, 6MWT | 98,44% |
| Angelucci et al. [94] | The NET-F index, HROS1% quantile, and average number of Fitbit step count correlated with 6MWD with 0.68 ($P < .001$), −0.39 ($P = .04$), and 0.59 ($P < .001$). The conditional probability of a 6MWT result $<$ 350 m if the mean number of steps is $<$ 10,000 is 0.53. The conditional probability of a 6MWT result $<$ 350 m if the maximum number of steps is $<$ 10,000 is 0.75. The ROC curve with NET-F index has an AUC of 0.80, and the optimal threshold, with a TPR of 0.88 and FPR of 0.43, is 7.78. This value can significantly discriminate the 6MWT walked distance ($P = .01$). | Wrist | IMU, PPG | 1 week | 6MWT | 82,76% |
| Moy et al. [52] | Steps per day were significantly associated with 6MWT distance (10 [4.17] Unadjusted Coefficient [95% CI]). | Shoes | IMU | 2 weeks | 6MWT | |
| Löppönen et al. [60] | Laboratory-based STS capacity was associated with the free-living mean STS performance ($r = 0.52$, $P < 0.001$) and free-living maximal STS performance ($r = 0.65$, $P < 0.001$). However, capacity and performance are not interchangeable but rather provide complementary information. Older and low-functioning individuals seemed to perform free-living STS movements at a higher percentage of their maximal capacity compared with younger and high-functioning individuals. | Upper thigh | IMU | 1 week | SPPB, STS | |

*Continued on next page*

| Article | Found association | Device Position | Type of sensor | Monitoring duration | Considered tests | Participants adherence |
|---|---|---|---|---|---|---|
| Saporito et al. [69] | Remote TUG can estimate TUG values with a median absolute error of 1.4 s (1.3 s), mean absolute error of 2.1 s (1.7 s), and AUC-ROC of 0.89. A correlation between the standardized TUG and the estimated TUG was found (0.70), indicating that there was agreement between sensor-based estimation and the standardized test value. | Chest | IMU | 3 days | TUG | 94,09% |
| Zheng et al. [76] | Daily steps were correlated with all physical performance scores in MS ($\|r_s\|$ = 0.58–0.60), but only correlated with 6MWT in healthy controls ($r_s$ = 0.38), all $P < .01$. Significant associations were observed between Peak30CAD and Peak-1CAD with SPPB, T25FW, 6MWT, and TUG in both groups. The magnitudes were stronger in MS than in controls ($\|r_s\|$ = 0.56–0.68 vs. 0.35–0.62). | Waist | IMU | 1 week | SPPB, T25FW, TUG, 6MWT | 62,62% |
| Apple Watch [55] | The ROC curve (AUC 0.946) is the result of e6MWD to classify users with respect to a threshold of 360 meters. | Wrist | IMU, PPG | Not specified | 6MWT | |
| Wright et al. [61] | Correlation between the physical capacity and physical activity was investigated. Largest correlation was 0.46 (p=0.157). The contributors were 6MWD, walking duration of longer bouts and movement intensity | Lower back | IMU | 1 week | 6MWT, SPPB, Sway Test, TUG | 72,38% |
| Santos et al. [62] | When observing results from each test we verified that spending more time in sedentary behaviors, independently of MVPA, had a negative impact on upper and lower body strength, agility/dynamic balance, and lower flexibility. On the other hand, independently of sedentary time, a positive association was found between MVPA with aerobic endurance and upper flexibility | Waist | IMU | 4 days | 30STS, Arm curl, Chair sit-and-reach, Back scratch, 8-foot up-and-go, 6MWT | |
| Nørgaard et al. [85] | 6MWD correlated 0.3, 0.29, 0.42 with accelerometer mean counts/min, number of minutes with counts > 1000/min and number of minutes with counts > 2500/min, respectively. | Waist | IMU | 1 week | 6MWT, Watt-max test (WMT) | 87% |
| Nero et al. [84] | Motor impairment is negatively associated with PA in older adults with PD. Motor impairment, physical function, body mass index and dyskinesia contributed to the variance of total physical activity, explaining 34% of the variance. | Waist | IMU | 1 week | Mini-BESTest | 91% |
| Agarwal et al. [86] | IMU derived features obtained in classifying performance values in the first quartile, interquartile range, and the fourth quartile were 0.62, 0.53, and 0.51 for the 400mWT, 20mWT, and 5STS, respectively (metric: The Goodman-Kruskal Gamma statistic) | Waist | IMU | 1 week | 400meterWT, 20meterWT, 5STS | |
| Urbanek et al. [63] | Median walking acceleration was associated with physical function (0.332, p = .001), median cadence was associated with gait speed (0.378, p = .002), and usual-paced 400mWT (-0.799, p = .034) and fast-paced 400mWT (-0.378, p = .002). Daily walking time was associated with fast-paced 400mWT (-0.318, p = .004). | Waist | IMU | 1 week | 400meterWT, 4meterWT | |
| O'Brien et al. [64] | Steps per day had moderate positive correlations with predicted VO2max, pushups and STS (r=0.37–0.58). Sedentary activity was not correlated to any physical fitness measures nor, grip strength or vertical jump. | Waist | IMU | 1-2 weeks | 6MWT, Grip strength, 30STS | |
| Van Gameren et al. [65] | Weak correlations between the SPPB and sensor-based daily life gait quality and quantity. This is likely given by the large interindividual variation in daily life gait quality and quantity data in combination with the small interindividual variation in the SPPB score in our study population of older adults | Lower back | IMU | 1 week | SPPB | 85% |
| Mate et al. [77] | Compared with people with 6MWT distance of at least 600 m, people walking less than 500 m had approximately half the rate of walking bouts of 5 minutes or longer. | Thigh | IMU | 1 week | 6MWT | 99.31% |
| Karavirta et al. [66] | 6MWT speed and acceleration significantly correlated with absolute MVPA (0.57, 0.49); with daily average acceleration (0.55, 0.49). Only 6MWT acceleration significantly correlated with relative physical activity (-0.30). Regression analysis showed that 6MWT speed explained 22% of the variation in absolute MVPA (p < .001). Relative PA was not significantly explained by 6MWT speed. | Thigh | IMU | 1 week | 6MWT | 89.7% |
| Birnbaum et al. [91] | Symptom severity, disease duration, lower limb strength were not associated with PA metrics. There was a positive correlation between PA and 6MWD (0.387, p=0.03) but not with sedentary behaviour (−0.183, p=0.32). | Lower back | IMU | 1 week | 6MWT, Isometric Kneww Extension strength | 86.84% |

*Continued on next page*

| Article | Found association | Device Position | Type of sensor | Monitoring duration | Considered tests | Participants adherence |
|---|---|---|---|---|---|---|
| Katzan et al. [54] | Correlations at baseline of step count with 2MWT, balance and 4mWT at discharge were 0.44 (95% CI −0.14 to 0.75),−0.12 (95% CI −0.67 to 0.64), and 0.17 (95% CI −0.46 to 0.66), respectively. Correlations at Day 30 were 0.22 (95% CI −0.45 to 0.71), −0.30 (95% CI −0.83 to 0.41), and 0.21 (95% CI −0.48 to 0.76) respectively. | Wrist | IMU, PPG | 90 days | Standing Balance Test, 2MWT, 4meterWT | 53.3% |
| De Pontes et al. [67] | Significant positive correlation was found between total energy expenditure and walking speed (0.266; p = 0.047), 6MWD (0.424; p = 0.001) maximum handgrip strength (r 0.478; p<0.001). | Thigh | IMU | 1 week | 4meterWT, 6MWT, Hand Grip Strength Test, TUG | |
| Choudhury et al. [68] | STS performance was associated with sedentary behaviour, light-intensity physical activity and MVPA durations in the regression analysis after adjusting for total wear time. | Wrist | IMU | 1 week | Hand Grip Strength Test, 30STS | 90% |
| Karle et al. [78] | No correlation was found between the 2MWT cadence and free-living walking minutes, in which participants reached 100%, 90% 80% and 70% of 2MWT cadence. | Lower back | IMU | 1 week | 2MWT | |
| Nair et al. [92] | Step activity correlated significantly with 10mWT (-0.56) and with 6MWD (0.40). Results showed that two years prior to LoA, on average, these boys had 32% lower daily step count compared to age-matched boys with DMD who remained ambulatory for at least two more years (3219 vs 4468, p=0.02). | Waist | IMU | 1 week, 1 and 2 year follow up | 6MWT, 10meterWT | 44.88% (1 year), 22.83% (2 years) |
| Low et al. [95] | Daily step count and SPPB (0.56 p<0.001), and TUG (-0.61, p<0.001). Peak gait cadence and SPPB (0.66 p<0.001), and TUG (-0.72, p<0.001). Activity fragmentation and SPPB (-0.69 p<0.001), and TUG (0.71, p<0.001). Time at home, Distance travelled and geographic mobility did not correlate significantly (p<0.01) with SPPB and TUG. | Undefined, wrist | IMU, GNSS, Smartphone interaction | 1 month | SPPB, TUG | 83.87 |
| Bachman et al. [97] | Spearman correlation of submaximal VO2 with sedentary time (-0.32 p<0.05), step count (0.63. p<0.001), light activity (0.4, p<0.01), moderate-to-vigorous activity (0.51, p<0.001), time in stepping bouts>1 minute (0.54, p<0.001) and peak 30s cadence (0.33, p<0.05). Weighted median cadence in stepping bouts>1 minute was the only measure not associated with submaximal VO2 (0.08; P=.61). After accounting for the effects of demographics and cancer characteristics in a partial Spearman correlation framework, average daily step count was significantly correlated with submaximal VO2 (0.46; P=.002), as was time in moderate to vigorous activity (0.33; P=.03). | Thigh | IMU | 1 week | Balke Treadmill Test | 56.98% |
| Chaudru et al. [80] | Moderate but significant inverse correlations were found between the mean walking pain manifestations per day and the maximal walking time (Gardner) (-0.472, P = .027) and the pain-free walking time (Strandness) (-.543, P = .016), the maximal walking time (Strandness) (-.527, P = .017). However, these correlations should be interpreted with caution considering the low number of walking pain manifestations experienced by the PAD participants. | Undefined, wrist | IMU, GNSS | 1 week | Gardner-Skinner treadmill protocol, Strandness treadmill protocol | 73% |
| Van de Port et al. [88] | The 10mWT was significantly associated with thenumber of steps a day and the step frequency (0.076, p:0.094*), TUG with step frequency (0.262 p=0.001). | Lower back | IMU | 1 week | TUG, 10meterWT | 85.71% |
| Rowlands et al. [81] | ISWT performance was positively correlated with PA volume (r=0.50, p<0.001) and absolute intensity (r=0.50, p<0.001), but negatively correlated with relative intensity (r=-0.13, p=0.025). Quantifying absolute and relative PA intensity of PA could improve enables personalisation of interventions. | Wrist | IMU | 1 week | ISWT | 81.79% |

Table 8. Risk of Bias Assessment

| | Confounding Bias | Selection Bias | Overall Bias |
|---|---|---|---|
| Wiedmann et al. [89] | + | - | - |
| Rens et al. [79] | - | - | - |
| Mueller et al. [87] | + | + | + |
| Ullrich et al. [82] | - | - | - |
| Sokas et al. [56] | X | - | X |
| Lyden et al. [93] | + | + | + |
| Sun et al. [70] | + | + | + |
| Crane et al. [57] | + | + | + |
| Straiton et al. [53] | - | X | X |
| Rekant et al. [58] | + | - | - |
| Zajac et al. [8] | + | + | + |
| Schütz et al. [59] | + | - | - |
| Bouça-Machado et al. [83] | + | - | - |
| Silva et al. [51] | - | X | X |
| Zhai et al. [71] | - | + | - |
| Block et al. [72] | + | + | + |
| Zhuparris et al. [90] | + | - | - |
| Block et al. [73] | + | + | + |
| Supratak et al. [74] | + | - | - |
| Motl et al. [75] | + | - | - |
| Angelucci et al. [94] | - | + | - |
| Moy et al. [52] | - | X | X |
| Löppönen et al. [60] | + | + | + |
| Saporito et al. [69] | - | - | - |
| Zheng et al. [76] | + | - | - |
| Apple Watch [55] | + | X | X |
| Wright et al. [61] | - | - | - |
| Santos et al. [62] | + | - | - |
| Nørgaard et al. [85] | + | + | + |
| Nero et al. [84] | - | - | - |
| Agarwal et al. [86] | - | + | - |
| Urbanek et al. [63] | + | - | - |

*Continued on next page*

**Table 8 – continued from previous page**

| | Confounding bias | Selection Bias | Overall Bias |
|---|---|---|---|
| O'Brien et al. [64] | + | - | - |
| Van Gameren et al. [65] | - | - | - |
| Mate et al. [77] | + | - | - |
| Karavirta et al. [66] | + | - | - |
| Birnbaum et al. [91] | - | - | - |
| Katzan et al. [54] | - | X | X |
| De Pontes et al. [67] | + | - | - |
| Choudhury et al. [68] | + | - | - |
| Karle et al. [78] | - | - | - |
| Nair et al. [92] | + | - | - |
| Low et al. [95] | + | - | - |
| Bachman et al. [97] | + | - | - |
| Chaudru et al. [80] | + | + | + |
| Van de Port et al. [88] | - | + | - |
| Rowlands et al. [81] | + | - | - |

*Key:* + *= Low Risk;* - *= Moderate Risk;* X *= Serious Risk.*