Forensic Science International 384 (2026) 112901

Contents lists available at ScienceDirect

Forensic Science International

journal homepage: www.elsevier.com/locate/forsciint

Review article

Is my smartwatch a valid witness? A systematic review and meta-analysis

Marcello Sicbaldi [a], Laura Bartoli [b,*], Jose Albites-Sanabria [a], Ilaria D'Ascanio [a], Alessandro Silvani [c], Lorenzo Chiari [a], Alberto Camon [b], Luca Palmerini [a]

[a] Department of Electrical, Electronic, and Information Engineering "Guglielmo Marconi", University of Bologna, Bologna 40136, Italy
[b] Department of Legal Studies, University of Bologna, Bologna 40126, Italy
[c] Department of Biomedical and Neuromotor Sciences, University of Bologna, Bologna 40126, Italy

ARTICLE INFO

Keywords:
Smartwatch
Wearable
Validity
Heart rate
Sleep
Forensics

ABSTRACT

Wearable devices are being increasingly used not only in traditional research fields (e.g. clinical, health) but also in innovative ones such as digital forensics. Modern commercial smartwatches, equipped with sensors like accelerometers, gyroscopes, and photoplethysmographs, can record activity, heart rate, sleep, and other physiological parameters that may support forensic investigations, given that their data is valid. To perform a first step in studying such validity, this article presents a systematic review and meta-analysis summarizing the available studies validating data collected from smartwatches, focusing on their potential use in forensic investigations. The review examined studies evaluating smartwatches from four major brands, comparing their outputs, such as activity classification, step count, distance, heart rate, energy expenditure, and sleep metrics, against established gold-standard measures. We found that not all brands of smartwatches and not all outcomes were equally studied, with varying results in terms of accuracy, protocols, and reporting of results. Heart rate resulted the most studied (and most accurate) measure. Most studies validated smartwatches in healthy populations and performed validation in laboratory settings. Overall, while smartwatches show promise for enhancing digital forensic analyses, their current validation evidence is heterogeneous. The findings highlight the need for more extensive and standardized validation efforts, larger and more diverse study populations, and improved transparency from manufacturers. These conclusions are not only relevant for forensic applications but also extend to broader domains, such as clinical and health research, where the reliability of wearable-derived data is equally critical.

1. Introduction

In the world of digital forensics, the analysis of wearable devices, or *wearable forensics*, is gaining increasing importance. With the growing use of devices such as smartwatches, fitness trackers, and smart glasses, these tools not only monitor our health and fitness activities but can also provide useful information for forensic investigations. Commercial smartwatches, which are now commonly worn by people of all ages, have integrated sensors such as accelerometers, gyroscopes, and photoplethysmographs (PPG), and may be equipped with a global positioning system (GPS). The variables measured by these sensors provide information on activity (number of steps, distance walked), sleep, and heart rate, offering a new landscape of information that may prove useful in solving complex cases. For example, this information may be leveraged, together with other sources of information, to verify alibis and to track movements, activities, and health condition of suspects and victims.

Case law is already filled with such examples especially in the U.S., where wearable technologies have been spreading earlier than anywhere else. Fitbit data have already been used on multiple occasions to disprove or make the prosecution's case. For instance, in a 2015 action for personal injury, Ms. Risley alleged to have been sexually assaulted while sleeping. The prosecutor was able to obtain the data of her Strava, which showed unequivocally that she had been walking around at the same time she was claiming to be sleeping. The State charged her with filing a false police report, and she plead guilty (Commonwealth of Pennsylvania v. Risley, No. CP-36-CR-0002937–2015).

Two years later, Mr. Dabate stated that a stranger broke into his home, assaulted him and killed his wife. No one saw the incident; the police searched the premises for traces and found Ms. Debate's fitness tracker, whose data sharply contradicted his story. The prosecution produced the data to show that the husband's tale was incompatible

* Corresponding author.
E-mail address: laura.bartoli2@unibo.it (L. Bartoli).

https://doi.org/10.1016/j.forsciint.2026.112901
Received 29 November 2025; Received in revised form 5 February 2026; Accepted 2 March 2026
Available online 6 March 2026

with what was recorded by the wearable device. He was later found guilty (State v. Dabate, Tolland Superior Court, Connecticut, Docket No: TTD -CR17–0110576-T).

Similar cases have occurred in Germany, U.K., Australia, and in Italy: the Assize Court of Bologna recently convicted a man for murder basing the judgment on circumstantial evidence and on the data extracted from the app 'iHealth', which consolidated the records of the iPhone, the iWatch, and the Garmin of the defendant (Assize Court of Bologna, 16 October 2024, n. 4; confirmed by the Assize Appellate Court of Bologna, 9 January 2026).

Among the challenges and limits related to the use of these technologies in the forensic field, there are: i) validity and accuracy of the extracted data, ii) privacy, personal data protection, and access to encrypted information; iii) data authenticity (possibility of manipulation of data). This article will focus on the first of these points, collecting and summarizing from the literature the available results on validity and accuracy of data collected from the most common smartwatches. This has not been extensively explored yet, especially in forensic contexts. Adding to the complexity of the topic, commercial smartwatches do not commonly store raw data (e.g. of acceleration signals, PPG signals…), but only processed outcomes such as steps and heart rate. Vendors do not provide information on their proprietary algorithms used to obtain those processed outcomes or on how they validated them, making the task more difficult. Moreover, vendors may modify their proprietary processing algorithms without prior notice, with potential long-term reproducibility issues. A recent work [1] proposed a forensics framework to analyze smartwatch health data and reported that the most commonly available health outcomes from such devices were steps, heart rate, heart rate variability, energy expenditure, distance, sleep, and activity recognition.

In the digital forensic context, the smartwatch sub-topic is relatively new, and few studies on smartwatches exist in general. Very few of them focus on the validation and the accuracy of the extracted outcomes. In [2], the comparison between two smartwatches, worn at the same time, showed differences among them, but did not use a gold standard (which would provide a reference to establish validity).

In other contexts, such as sport, fitness, and clinical research, several studies have analysed the validity of outcomes from commercial smartwatches, focusing on metrics that are also of interest from a forensic perspective, especially steps, heart rate, and energy expenditure. The first two metrics have been proven useful in multiple cases (see Schlusche et al. [3] for an example); the latter has been reportedly used in a rape and murder case to support the interpretation that the defendant performed a physically demanding activity consistent with carrying the unconscious victim to the riverbank, where she was ultimately found dead [4]. Different reviews have addressed the accuracy of fitness trackers. Germini et al. [5] found that the Fitbit Charge's step count accuracy has a mean absolute percentage error (MAPE) of less than 25%, whereas its energy expenditure measurements have a MAPE greater than 30%. Fuller et al. [6] compared smartwatches from several brands, noting that i) Fitbit, Apple Watch, and Samsung measured steps accurately in laboratory settings; ii) Apple and Garmin were the most accurate for heart rate measurement, whereas Fitbit tended toward underestimation; iii) no smartwatch was accurate for energy expenditure. Chevance et al. [7] focused on Fitbit devices, reporting that they underestimate heart rate (-2.99 bpm), energy expenditure (-2.77 kcal/-min), and step rate (-3.11 steps/min). Evenson and Spade [8] reported MAPE values to quantify the accuracy of Garmin devices. Lu et al. [9] characterized remote cardiovascular monitoring technologies, highlighting the use of PPG in more than half of the cases. For sleep tracking, Lee et al. [10] performed a prospective multicenter validation study of 11 wearable, nearable, and airable consumer sleep trackers, highlighting substantial performance variability, with macro F1 score ranging from 0.26 to 0.69. Schyvens et al. [11] evaluated three recent sleep trackers (Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP), concluding that all devices can benefit from further improvement in the assessment of specific sleep stages.

Other examples are present if the search is extended to portable digital devices such as smartphones. In the forensic area, Van Zandwijk et al. [12] validated the identification of ascending and descending stairs using an iPhone. Regarding health and clinical applications, Werner et al. [13] performed validation of walking outcomes (steps, walking speed, stride length) obtained with iPhones (a study with a similar aim was reported in Apple [14] by the manufacturer itself).

Most reviews identified the following aspects as the ones where improvement should be sought: reliable translation to real-world scenarios, inclusion of diverse demographics, and the use of standardized processes for data collection and device validation. The need of gold standard systems that are appropriate for the intended application of the sensors has also emerged as an important topic, together with the concept of "performance evaluation" for validation studies (i.e. to consider a wearable sensor as appropriate for research or clinical studies, one should not only consider whether a validation study is present but also the obtained performance) [15].

The aim of this study was to conduct a systematic review and meta-analysis of the performance, assessed against a gold standard (which is needed to establish validity and accuracy), of recent smartwatches from some of the most popular and common brands on the market. The selected smartwatches were the ones with a release date from 2019 onward and produced by four brands: Apple, Fitbit, Garmin, and Samsung. We focused on the information that is most commonly provided by smartwatches, and that could be useful in a forensic context, such as number of steps and distance walked, heart rate, heart rate variability (providing information related to stress) [16] energy expenditure, sleep, and physical activity (e.g., running).

Previous literature reviews have examined the validity of consumer wearable devices, typically focusing on specific products or brands, or on limited subsets of measured outcomes (e.g., step count, heart rate, energy expenditure, or sleep). However, relatively few reviews adopt a perspective aligned with forensic and investigative contexts. In this work, we therefore prioritize widely adopted smartwatch ecosystems and recent device generations, reflecting those most likely to be present in opportunistic, non-controlled scenarios relevant to digital investigations.

Furthermore, only a limited number of prior reviews have included a quantitative synthesis of findings across studies. To our knowledge, the only work combining a systematic review with a meta-analysis in this domain is that of Chevance et al. [7], which was restricted to Fitbit devices and a narrow set of metrics (steps, energy expenditure, and heart rate), whereas the present review addresses a broader range of devices and outcomes of forensic relevance.

2. Methods

2.1. Systematic review

We performed our search on PubMed and Scopus datasets (see 2.1 and 2.2 for specific search strings) in the last 5 years. The search strings were designed to provide results on the following outcomes: activity classification (e.g., walking, stairs, cycling, running, driving a car…), steps, distance walked, sleep, heart rate, and heart rate variability. After removing duplicates, two independent reviewers selected the articles. The selected articles were then screened reading their full-text content for extraction of information.

2.1.1. PubMed search string

Below the components of the PubMed search string:

- #1 ("garmin" OR "fitbit" OR "apple watch" OR "samsung galaxy")
- #2 ("validity" OR "validation" OR "validated" OR "accuracy" OR "accurate" OR "reliability" OR "reliable" OR "performance evaluation")

- #3 (“step*” OR “floor*” OR “flight*” OR “activity” OR “sport” OR "exercise" OR ”stair*” OR “walking” OR "run" OR “running” OR “distance” OR “energy expenditure” OR "swimming" OR "swim" OR "cycling" OR "bike" or "bicycle" OR "driving" OR "drive" OR "car" OR “gait”)
- # 4 (“heart rate”)
- # 5 (“sleep”)
- #6 (“position” OR “location” OR "GPS").

The following is the complete search string, using the components mentioned above:

- #1 AND #2 AND (#3 OR #4 OR #5 OR #6) AND (2020:2025[pdat])

2.1.2. Scopus search string

The following is the Scopus search string:

TITLE-ABS-KEY(("garmin" OR "fitbit" OR "apple watch" OR "samsung galaxy") AND ("validity" OR "validation" OR "validated" OR "accuracy" OR "accurate" OR "reliability" OR "reliable" OR “performance evaluation”) AND ("step*" OR "floor*" OR "flight*" OR "activity" OR "sport" OR "exercise" OR "stair*" OR "walking" OR "run" OR "running" OR "distance" OR "energy expenditure" OR "swimming" OR "swim" OR "cycling" OR "bike" or "bicycle" OR "driving" OR "drive" OR "car" OR “gait” OR “heart rate” OR “sleep” OR "GPS" OR "position" OR "location" OR "GPS")) AND PUBYEAR $>$ 2020 AND PUBYEAR $<$ 2025

2.1.3. Inclusion criteria

The following inclusion criteria were considered:

- Journal or conference paper in English
- Smartwatches released in or after 2019 from Apple, Fitbit, Garmin, and Samsung,
- Focus on activity classification (e.g., walking, stairs, cycling, running, driving a car…), steps, distance walked, energy expenditure, sleep, heart rate, and heart rate variability.
- Validation or performance evaluation against a recognized gold standard (also reported as criterion or reference system):
- Studies with the availability of a gold standard. Gold standard measures were pre-specified a priori for each outcome based on established practice in validation studies:
  - o For activity classification: video or direct observation
  - o For steps: Stereophotogrammetry, instrumented mat/treadmill, video, hand tally counter
  - o For distance walked: a-priori known path, podometer (wheel meter).
  - o For sleep: polysomnography
  - o For energy expenditure: Indirect calorimetry
  - o For heart rate and heart rate variability: standard electrocardiography or wearable Polar H10 system [17]

2.1.4. Exclusion criteria

The following exclusion criteria were considered:

- Review articles
- Less than 5 subjects examined
- Subjects under 14 years of age (i.e., minimum age of criminal responsibility in many European countries, including Italy)
- Numerical values of the performance metrics not available in published tables or text (e.g., availability of graphical Bland-Altman plots only)

2.1.5. Extraction of information

The following information was extracted from each article at the full-text screening phase:

Title, year, smartwatch model(s) considered, clinical information and demographics of the participants to the study, number of subjects and samples in the analysis, protocol information, gold-standard (criterion) measurement, and validity performance measures. Regarding validity performance measures, we extracted bias (average of the differences between the tested device measurement and the corresponding gold-standard measurement) and limits of agreement (95% confidence interval of the difference between the device measurement and its gold standard measurement) for steps, heart rate, heart rate variability, energy expenditure, distance, and total sleep time. For sleep classification (wake vs. sleep epochs), we considered sensitivity and specificity in detecting sleep. For activity recognition, we considered sensitivity, specificity, and accuracy.

Finally, the modality of data extraction from the smartwatch (e.g., software/app) was also collected.

2.2. Meta-analysis

The extracted information was analyzed in a meta-analysis following the framework utilized in Chevance et al. [7]. We performed the meta-analysis divided for each brand, performing it for a specific outcome only if data from at least two studies were available. Differences with respect to the methodology proposed in [7] are reported in (Multimedia Appendix 1). We performed the meta-analysis on continuous outcomes, using the Bland-Altmann metrics of agreement: bias and limits of agreement (LoA). For non-continuous outcomes (activity and sleep epoch classification) we only reported results without performing meta-analysis.

2.3. Risk of bias, outliers, sensitivity, and subgroup analysis

Following [7], which was a customized version of the COSMIN checklist [18], we identified 4 points to assess study quality: 1) sample size of at least 50 subjects (1 point), 2) presence of peer-review (1 point); 3) device firmware details (i.e. version) provided (1 point); 4) validation of only one device on the wrist at the same time (1 point). The latter point is to assign higher quality to studies where only one smartwatch was tested at a specific time point, so that the smartwatch could be worn in the most proper and comfortable position. The total points can range from low (1) to high (4) scores.

To evaluate the robustness of our analysis we tested the sensitivity to bias (by re-performing the analysis excluding studies with a high risk (a score $\leq$1) of bias.

Then, to test the robustness with respect to outliers, we re-performed the meta-analysis after discarding outliers using the Grubbs test, as in [7].

Finally, to provide more detailed insights on the results, we performed subgroup meta-analyses for the following variables: (1) characteristics of the participants, including the presence of health conditions and age ($<$65 and 65 $+$); (2) device model; (3) type of activity (e.g., resting, walking, and cycling); (4) intensity (i.e., differences between light and moderate to vigorous intensity activities); and (5) study quality. Subgroup analyses were conducted when at least four comparisons between the device measures and criterion measures were available. If in a study there was more than one test for the same subjects (e.g. heart rate recording during walking and running), those tests were considered as different comparisons.

All analyses were conducted using the R statistical program (version 4.4.2; R Foundation for Statistical Computing). The R code was adapted from the study by [7], which was itself adapted from [19]).

3. Results

3.1. Review

In (Fig. 1), the PRISMA flow chart [20] and related results in abstract and full-text screening are reported. From a total of 795 articles, 651 remained after duplicate removal. After review of the abstract from two

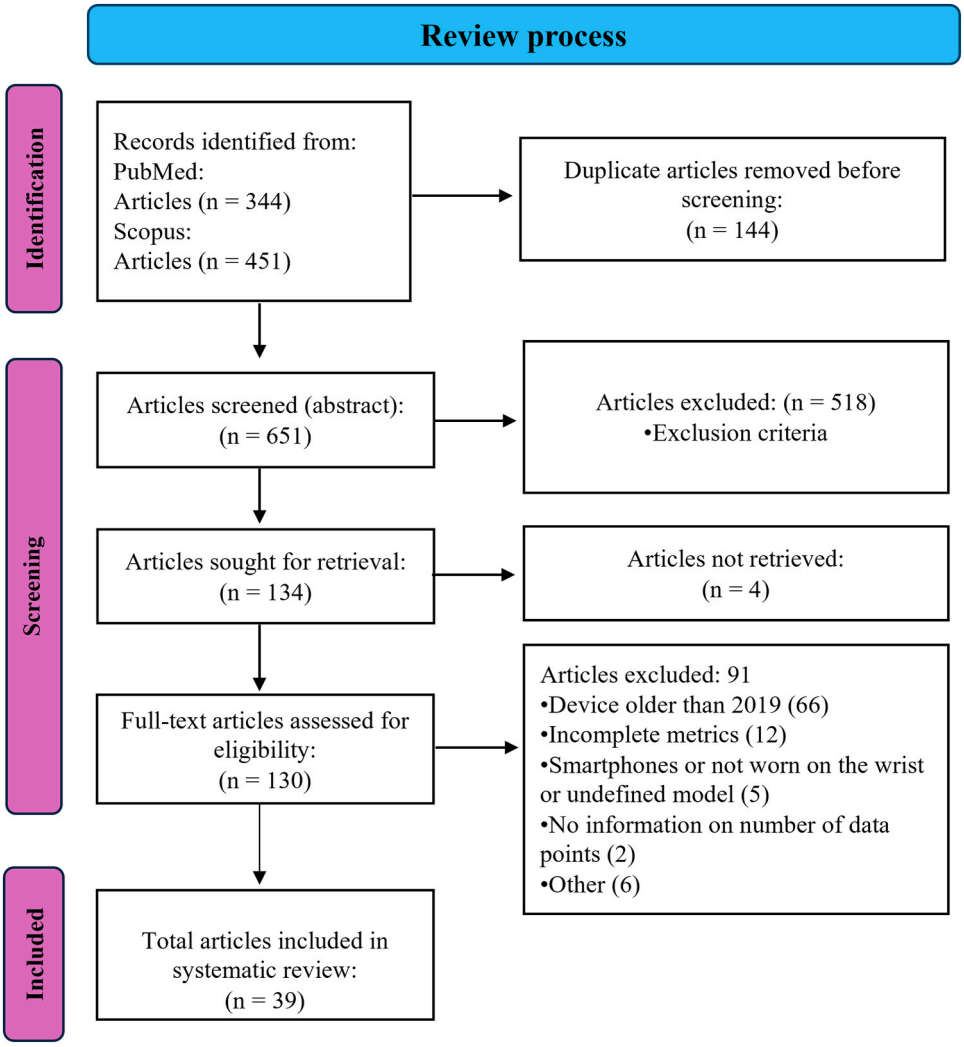

Fig. 1. Flow diagram of the review process.

independent raters, 134 articles were selected. For 130 articles it was possible to obtain the full-text and those proceeded to full-text screening. Here, 91 articles were discarded, mainly (66/91) because they considered smartwatches with a release date earlier than 2019. In the end, 39 articles were selected for information extraction and meta-analysis.

(Fig. 2) presents the distribution of various brands and models that were considered in the selected articles.

(Table 1) presents the details of the selected articles, showing the main related information: participants characteristics, devices analyzed, validated outcomes (steps, heart rate.), gold standard (criterion), setting (Lab or Field, depending if the test was performed in the laboratory or hospital (Lab) or outside (Field)), protocol, and the information whether a single device was worn on the wrist in the experiment (Table 1-3).

In (Multimedia Appendix 2), further details on each study are presented, divided by brand, with the value of the performance metrics and with the information needed to assess study quality.

Considering the selected articles, the majority validated Fitbit and Apple brands (see Fig. 2). Samsung watches were less studied. Garmin had the largest range of studied models, with 7 different models in 12 studies.

Most of the 39 identified articles validated smartwatches in healthy populations (66.7%), with a mean age of 14.7–74 years old. The number of studied subjects varied from 9 to 263. Clinical populations included Parkinson's disease, stroke, and sleep disorders (see Table 1). Most studies (23.6%) tested heart rate, followed by sleep epoch classification (23.1%), total sleep time (15.4%), energy expenditure (12.8%, expressed either in MET or in kcal), steps (10.3%), and heart rate variability (5.1%). Studies could test more than one outcome at once. No study tested activity recognition (such as identification of running/cycling/stairs…). The selected position on the wrist was highly variable and mostly either randomly assigned (48.7%) or at the non-dominant wrist (23.1%). Most studies (32 out of 39) were performed in a laboratory setting, 6 in the field, and one both in the laboratory and in the field. Only 2 studies studied real-world scenarios (coded as a combination of field as setting and daily activities in Table 1). The Fitzpatrick scale for skin tone was reported only in 5 studies that addressed heart rate and heart rate variability through PPG sensors (Multimedia Appendix 2). Bias was not always reported as in the conventional difference between measurements with the tested device and the gold standard: 7 studies out of 39 reported differences between the gold standard and the tested device. Finally, all studies satisfied at least one quality criterion; 6 studies satisfied only one, 24 satisfied 2 out of 4, 8 studies 3 out of 4, and one study satisfied 2 or 3 criteria depending on the target population. None of the studies satisfied all 4 quality criteria.

3.2. *Meta-analyses*

(Table 2), and (Figs. 3,4,5,6) present the results of the meta-analyses assessing the accuracy of commercial smartwatches.

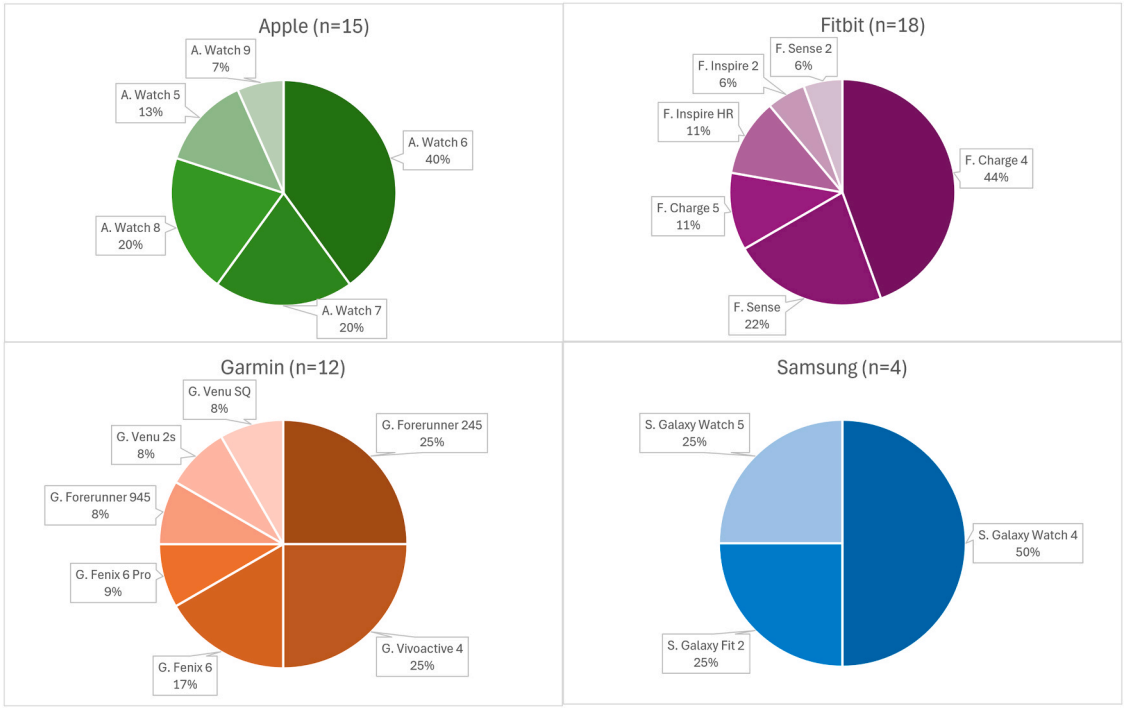

Fig. 2. Models and brands of the smartwatch analysed in the selected articles.

3.2.1. Heart rate

Apple, Garmin, and Samsung smartwatches showed minimal discrepancies with respect to the actual values measured with the reference, with a bias (average difference with respect to the reference value reported by the gold standards) close to zero, and moderate variability, with limits of agreement approximately within ± 10 bpm ((Fig. 3), left). Apple and Garmin showed a slight underestimation of heart rate with respect to the gold standard, with an average bias (standard deviation, SD) of −0.7 bpm (SD 4.1 bpm) and −0.9 bpm (SD 4.3 bpm), respectively. Samsung showed a slight overestimation of heart rate, with a mean bias of 0.5 bpm (SD 5.7 bpm). Fitbit devices exhibited a more substantial underestimation of heart rate, with a mean bias of −3.4 bpm (SD 8.0 bpm) and wide limits of agreement from −21–14 bpm, indicating lower consistency. Garmin, Apple, and Fitbit had a high number of available comparisons between the device and the gold standard within studies (between 27 and 35), while Samsung was less studied (6).

3.2.2. Heart rate variability

It was not possible to perform meta-analysis on heart rate variability, since only one study was present, both for the Fitbit and the Apple brands. For the former, the root mean square of successive differences (RMSSD) in milliseconds (ms) was evaluated during the night in healthy subjects [33], with an underestimation of −3 ms in healthy subjects and limits of agreement between −22 ms and 15 ms. In subjects with chronic obstructive pulmonary disease, an underestimation of −7 ms with limits of agreement of −30 ms and −16 ms was found. For comparison, the average (SD) values for healthy subjects from the study were reported as 33 (21) ms. In the latter study, the mean difference in heart rate variability (evaluated through SDNN, standard deviation of normal-to-normal inter-beat intervals) between Apple Watch and the reference standard was 8.31 ms (overestimation by Apple Watch), with a confidence interval between 5.6 and 11 ms, and a reference average value of 85 ms.

3.2.3. Energy expenditure

Apple and Fitbit were the only brands with enough studies for the meta-analysis of energy expenditure. Available comparisons of EE were much lower than the ones available with heart rate, with Apple having the most (n = 9).

Apple watches underestimated energy expenditure, evaluated in kCal, by about −9.68 kcal (SD 12.23), with large variability as demonstrated by their wide limits of agreement ranging from −45–26 kcal ((Fig. 4), left). For a scale reference, Hajj-Boutros et al. [29] reported energy expenditures of 14, 45, and 117 kcal, due to 10 min of resting, walking, and running, respectively.

Fitbit overestimated energy expenditure, evaluated with METs, by two METs on average, with a large confidence interval ((Fig. 4), right). For a scale reference, light activity (casual walking, doing household chores, or activities of daily living) is usually associated with 1.6–2.9 METs, moderate activity (e.g., brisk walking) with 3–5.9 METs, and vigorous activities (e.g., sport) with values equal to or greater than 6 METs.

3.2.4. Steps

Fitbit and Garmin were the only brands with at least two studies considering the number of detected steps. Fitbit (with 3 available comparisons) underestimated steps by an important amount (-138 steps), with a large 95% confidence interval (-343.94–68.39 steps), as shown in (Fig. 5). The reference steps in the different available comparisons ranged between 400 and 800 steps performed on average by each subject. Garmin, on the other hand, with 6 available comparisons, showed only a slight underestimation of the reference values (-4.8 steps) and a 95% confidence interval between −60 and 50 steps. The reference steps in the different available comparisons ranged between 0 and 800 steps performed on average by each subject.

3.2.5. Distance

It was not possible to perform meta-analysis on results concerning distance, since only one study was retrieved, considering one Fitbit and one Garmin device. Both smartwatches underestimated the distance walked in two tests, with a bias between −54 and −144 m, with a reference value of distance being between 649 and 679 m for each subject.

Table 1
Selected articles.

Study	Participants (cohort; N participants; age mean ± std; % female)	Device	Outcome	Gold standard	Wrist	Setting Protocol	Single device
Budig et al. [21]	Healthy N 36; age 36.1 ± 12.8; F 44.4%	G. Forerunner 245	HR	Polar	Random	Lab Mix (cycling, running, swimming, walking)	no
Budig et al. [22]	Sleep disorders N 49; age 55.01 ± 10.2; F 63.3%	G. Forerunner 945	Sleep (TST)	PSG	Random	Lab Sleeping	no
Ceugniez et al. [23]	Healthy N 21.1; age 21.1 ± 5.8; F 44.0%	F. Charge 4	HR	Polar	Non-dominant	Field Badminton, cycling, orienteering running, soccer, tennis	yes
Colonna et al. [24]	Parkinson's disease N 11; age 65.3 ± 5.8; F 45.5%	F. Charge 4	HR	Polar	Side less affected by Parkinson's disease	Lab Resting, maximal exercise	yes
Cosoli et al. [25]	Healthy N 10; age 17.3 ± 3.3; F NA	G. Venu SQ	HR	Polar	Right	Lab Resting	yes
Dong et al. [26]	Sleep disorders N 11; age 65.3 ± 5.8; F 45.5%	F. Charge 4	Sleep (epoch) Sleep (TST)	PSG	Non-dominant	Lab Sleeping	yes
Giggins et al. [27]	Healthy N 10; age 30.4 ± 8.0; F 50.0%	A. Watch 5	HR	Polar	Left	Field Daily activities	yes
Grandner et al. [28]	Healthy N 36; age 23.4 ± 3.8; F 44.4%	F. Charge 4	Sleep (epoch)	PSG	Left	Lab Sleeping	no
Hajj-Boutros et al. [29]	Healthy N 60; age 24.9 ± 3.0; F 50.0%	A. Watch 6 F. Sense	HR EE	Polar MetaMax 3B	Dominant	Lab Resting, walking/running (treadmill), resistance exercises, cycling	no
Helmer et al. [30]	Postoperative N 112; age 68.0 ± 11.9; F 37.5%	A. Watch 7 F. Sense G. Fenix 6 Pro	HR	ECG	Random	Lab Resting	no
Helmer et al. [31]	Postoperative N 10; age 63.0 ± 13.5; F 70.0%	A. Watch 7 G. Fenix 6 Pro	HR	ECG	Random	Lab Mix (hospital stay)	yes
Hermand et al. [32]	Healthy N 10; age 32.5 ± 9.4; F 50.0%	G. Forerunner 245	HR	ECG	Random	Lab Cycling	yes
Hermans et al. [33]	Healthy N 20; age 68.0 ± 7.0; F 48.2% COPD N 25; age 70.0 ± 7.0; F 20.0%	F. Charge 4	HR HRV	Polar	Non-dominant	Field Resting	yes
Ho et al. [34]	Healthy N 30; age 29.3 ± 3.6; F 50.0%	A. Watch 6	HR	ECG	Left	Lab Cycling	yes
Jachymek et al. [35]	Healthy N 31; age 28.0; female 32.3%	F. Charge 4	HR	ECG	Left	Lab Running (treadmill)	yes
Jamieson et al. [36]	Healthy N 16; age 27.0 ± 2.2; F 56.0%	F. Charge 4 G. Vivoactive 4	HR Steps	ECG Hand tally counter	Random	Field HR: Resting, walking, alternate stepping, chair rising, recovery; Steps: walking	yes
Kastelic et al. [37]	Healthy N 28; age 74.0 ± 5.0; F 53.6%	G. Vivoactive 4	Steps	video	Non-dominant	Lab Walking	no
Khushhal et al. [38]	Chronic disease; N 41; age 51.0 ± 7.0; F 50.0%	A. Watch 8	HR	Polar	Left	Lab Cycling	yes
Kim et al. [39]	Cardiovascular disease N 78; age 59.9 ± 9.4; F 16.7%	S. Galaxy Fit 2	HR	ECG	Random	Lab Running (treadmill)	yes
Kim et al. [40]	Coronary artery disease; N 44; age 60.9 ± 7.9; F 18.2%	A. Watch 7 S Galaxy Watch 4	HR	ECG	Random	Lab Running (treadmill)	yes
Le et al. [41]	Healthy N 20; age 23.0 ± 2.2; F 50.0%	A. Watch 6 G. Fenix 6	EE kcal	COSMED K5	Random	Field Walking	no
Lee et al. [42]	Sleep disorders N 26; age 43.6 ± 14.1; F 48.0%	A. Watch 8 F. Inspire HR S. Galaxy Watch 5	Sleep (epoch) Sleep (TST)	PSG	Random	Lab Sleeping	no

(continued on next page)

Table 1 (*continued*)

Study	Participants (cohort; N participants; age mean ± std; % female)	Device	Outcome	Gold standard	Wrist	Setting Protocol	Single device
Li et al. [43]	Postoperative N 45; age 52.2 ± 13.6; F 40.0%	F. Inspire HR	Steps	Hand tally counter	Non-dominant	Lab Walking	no
Lim et al. [44]	Healthy N 9; age 39.0 ± 8.0; F 55.6%	F. Inspire 2	Sleep (epoch) Sleep (TST)	PSG PSG	Random	Lab Sleeping	yes
Lima et al. [45]	Healthy N 14; age 28.6 ± 5.6; F 50.0%	S. Galaxy Watch 4	HR	Polar	Random	Lab Cycling	yes
Meza et al. [46]	Stroke N 70; age 79.4 ± 10.1; F 63.0%	F. Charge 5	HR	ECG	Non-dominant	Lab Resting	no
Miller et al. [47]	Healthy N 53; age 25.4 ± 5.9; F 49.6%	A. Watch 6 G. Forerunner 245	Sleep (epoch) Sleep (TST) HR	PSG (sleep) ECG (HR)	Random	Lab Sleeping	no
Neudorfer et al. [48]	Healthy N 32; age 59.1 ± 10.1; F 34.4%	G. Venu 2 s	HR	ECG	Random	Lab Field Cycling, daily living Daily living	yes
Ogasawara et al. [49]	Psychiatric disorders N 52; age 48.1 ± 17.2; F 52.0%	F. Sense	Sleep (epoch) Sleep (TST)	PSG	Non-dominant	Lab Sleeping	no
O'Grady et al. [50]	Healthy N 39; age 24.6 ± 8.2; F 49.6%	A. Watch 9	HR HRV	Polar	Random	Field Resting	yes
Ong et al. [51]	Healthy N 60; age 38.5 ± 15.1 F 56.1%	F. Sense	Sleep (epoch) Sleep (TST)	PSG	Non-dominant	Lab Sleeping	no
Robbins et al. [52]	Healthy N 33; age 32.9; F 57.1%	A. Watch 8 F. Sense 2	Sleep (epoch) Sleep (TST)	PSG	Random	Lab Sleeping	no
Simon-Vicente et al. [53]	Huntington's disease N 14; age 55.7 ± 11.4; F 50.0%	F. Charge 4	EE MET	Medisoft Ergo Card	Dominant	Lab Walking (treadmill)	yes
Støve et al. [54]	Healthy N 29; age 24.5 ± 4.0; F 55.0%	A. Watch 6	HR	Polar	Random	Lab Resistance exercises, recovery	yes
Stutz et al. [55]	Healthy N 60; age 64.0; F 48.0%	F. Charge 5	EE MET	Oxycon, Metamax 3B	Dominant	Lab Mix (ADLs, cycling, physical exercises, walking, resting, running, stairs)	no
Sun et al. [56]	Healthy N 11; age 22.5 ± 1.8; F 55.0%	A. Watch 6	EE kcal	MetaMax 3B	Random	Lab Running (treadmill), running	no
Theurl et al. [57]	Myocardial infarction N 104; age 59.0 ± 8.9; F 17.0% Stroke N 129; age 67.0 ± 13.3; F 30.0% Healthy N 30; age 44.0 ± 23.7; F 47.0%	G. Vivoactive 4	HR	ECG	Right	Lab Resting	yes
Viciana et al. [58]	Healthy N 56; age 14.7 ± 1.7; F 48.2%	A. Watch 5	Steps	Video	Non-dominant	Lab Walking	no
Willoughby et al. [59]	Healthy N 60; age 39.6 ± 15.5 F 57.1%	F. Sense	Sleep (epoch) Sleep (TST)	PSG	Random	Lab Sleeping	no

N: number of participants, F: female; A.: Apple device, F.: Fitbit device, G.: Garmin device, S.: Samsung device; HR: heart rate, EE: energy expenditure, MET: metabolic equivalent task, TT: total sleep time; ECG: electrocardiogram, Polar: Polar H10, PSG: polysomnography; Lab: laboratory.

Table 2
Results of the meta-analysis.

Analyses	k comparisons	Bias, mean (SD)	τ	LoA	95% CI
Main Analyses					
HR (bpm)					
Apple	33	-0.62 (4.12)	0.25	-8.91–7.68	-12.08–10.84
Fitbit	27	-3.44 (8.01)	14.2	-21.14–14.26	-27.64–20.76
Garmin	35	-0.91 (4.31)	1.87	-9.95–8.12	-11.63–9.79
Samsung	6	0.52 (5.71)	0.89	-11.04–12.09	-17.48–18.52
EE (Kcal)					
Apple	9	-9.68 (12.23)	167.38	-45.29–25.91	-71.86–52.49
Fitbit	5	-9.37 (10.83)	250.77	-47.75–29.01	-92.06–73.32
Garmin	2	23.03 (30.33)	22.39	-38.36–84.43	-252.35 298.42
EE (MET)					
Fitbit	4	1.98 (1.84)	3.66	-3.32–7.30	-8.70–12.68
Steps					
Fitbit	2	-137.78 (89.11)	2684.15	-343.94–68.39	-584.02–308.47
Garmin	2	-4.76 (26.26)	59.77	-59.51–49.99	-84.57–75.05
TST (minutes)					
Apple	2	23.46 (29.68)	472.7	-50.1–97.04	-429.97–476.89
Fitbit	4	-4.91 (37.66)	48.51	-81.50–71.69	-106.05–96.24

k is the number of comparisons between the device and criterion measures available within studies.
Bias is the pooled estimate of mean differences calculated as Tested device – criterion (i.e., smartwatch – gold standard).
SD is the pooled standard deviation of differences.
τ is the variation in bias between studies.
LoA: limits of agreement: lower (upper) 95% limit of agreement was calculated from pooled estimates of bias and SD of differences with robust variance estimation.
95% CI: outer confidence bound for lower and upper 95% limit of agreement.
HR: heart rate; bpm: beats per minute.
EE: energy expenditure; kcal: kilocalories; MET: metabolic equivalent of task.
TST: total sleep time

Table 3
Data extraction modalities reported in included studies.

Brand	Extraction route reported in included studies	Exported / format (as reported)	Data type (raw vs processed)
Apple Watch	Third-party apps (Sleep Watch; Breathe) and Apple Health export	App-derived outputs; XML export from Health app including processed parameters (e.g., heart rate, steps, sleep); geolocation data when a workout is initiated	Processed outcomes (vital parameters, activity/sleep summaries); geolocation tied to workouts
Fitbit	Web platform download; Pulse Watch mobile application	Platform/app downloads (format not specified)	Processed outcomes (as reported)
Samsung	Manufacturer application or web portal	App/portal download (format not specified in paragraph)	Processed outcomes (as reported)
Garmin	Manufacturer application or web portal; also direct USB transfer to computer (in one study)	App/portal download (format not specified); USB-downloaded. fit file (reported in one study)	Processed outcomes; device file export (.fit) used for subsequent processing
Cross-brand (reported for energy expenditure)	In one comparison, energy expenditure was obtained from screenshots of the respective app	Screenshots	Processed outcomes only (displayed estimates)

3.2.6. Sleep

In the domain of sleep tracking (Fig. 6), Apple watches overestimated total sleep time by 23.5 min (SD 29.7 min). Conversely, Fitbit underestimated total sleep time by −4.9 min (SD 37.7 min). Confidence intervals were −50.1–97.0 min (Apple) and −81.5–71.7 min (Fitbit).

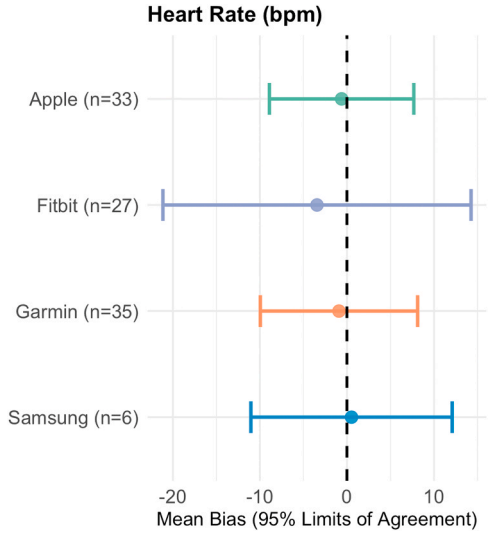

Fig. 3. Forest plots for the main analyses of heart rate. N indicates the number of available comparisons between the tested device and the gold standard. Bpm, beats per minute.

Nine studies addressed the discrimination of wakefulness from sleep (2-state classification). All devices showed high ability in correctly detecting sleep (high sensitivity), but lower ability in detecting wakefulness (low specificity). For Fitbit studies (8 comparisons), sensitivity ranged from 89% to 96%, with low variability, while specificity ranged from 13% to 78%, with substantial variability. For Apple watches (3 comparisons), sensitivity similarly ranged from a minimum of 96% to a maximum of 97%, with low variability, while specificity ranged from a minimum of 26% to a maximum of 52%, with substantial variability (Multimedia Appendix 2). Only one comparison was present for Samsung and Garmin watches, with similar results, with sensitivity of 91% and 98%, respectively, and specificity of 48% and 27%, respectively.

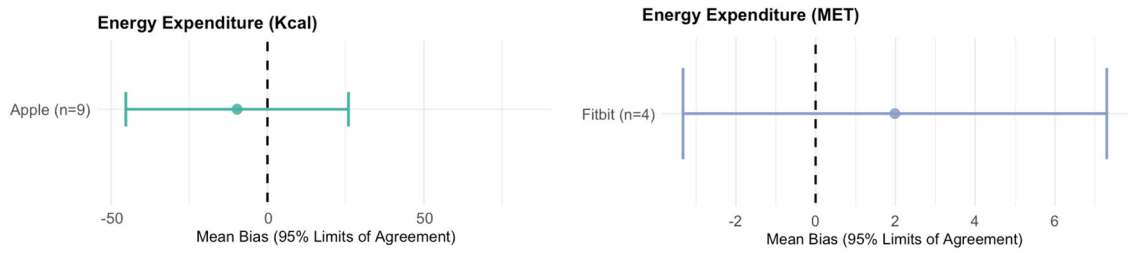

Fig. 4. Forest plots for the main analyses of energy expenditure, measured in kilocalories (kcal), on the left, and metabolic equivalent task (MET), on the right. N indicates the number of available comparisons between the tested device and the gold standard.

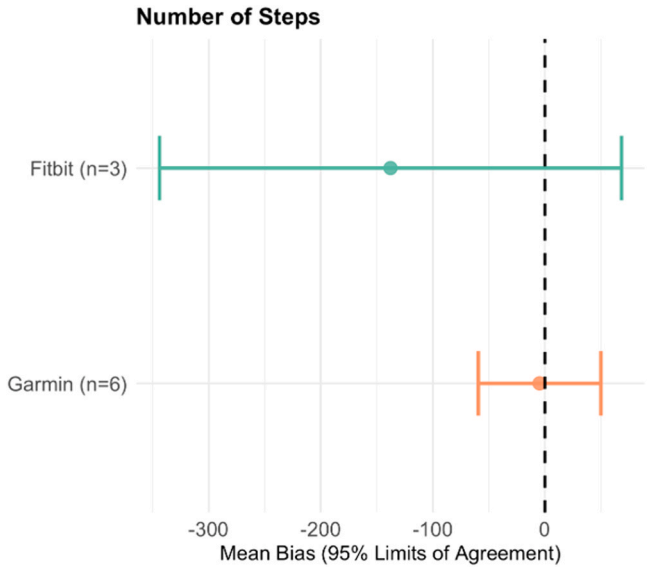

Fig. 5. Forest plots for the main analyses of steps. N indicates the number of available comparisons between the tested device and the gold standard.

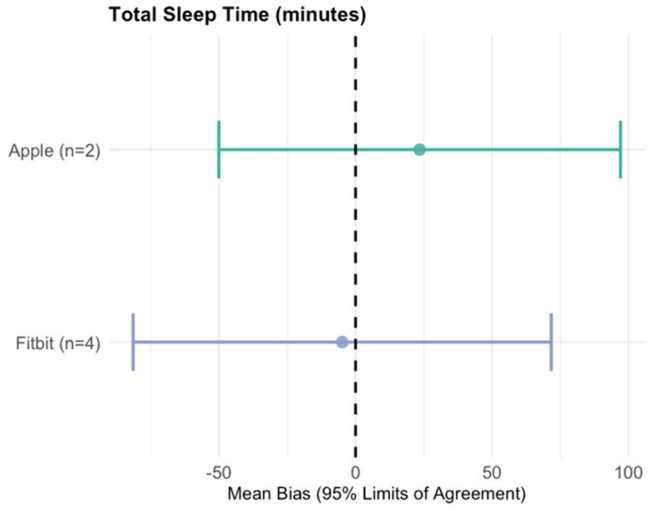

Fig. 6. Forest plots for the main analyses of distance (meters) on the left and total sleep time (minutes) on the right. N indicates the number of available comparisons between the tested device and the gold standard.

3.3. Secondary analyses

The results of the sensitivity analyses and the analyses after discarding the outliers are reported in (Multimedia Appendix 1).

No analysis on sensitivity to bias could be performed because i) in terms of heart rate measurements, which had the largest sample size, no study had low quality scores (all had quality > 1), and ii) for the other variables, even if some studies did get low quality scores, the number of studies available was too small.

Subgroup meta-analyses were performed for heart rate only, as no other variable had at least four comparisons between device measures and criterion measures. Overall, subgroup analyses by population (healthy subjects vs. clinical samples), age, exercise (rest vs. activity), model brands, context (laboratory vs. field recording), showed an underestimation of heart rate by the devices compared with criterion measures in most cases (Fig. 7), consistent with the main analysis without subgroups. Field measurements and measurements in clinical populations showed higher limits of agreement, as expected. Variability was also present in the performance of different models of smartwatches from the same brand.

3.4. Data extraction from smartwatches

In the current review, different researchers used different methodologies to acquire the data to perform the comparisons. Not all studies, however, reported the methodology.

For Apple Watch devices, we found that data were obtained through third-party applications such as Sleep Watch [47], which measures heart rate for each sleep period, and the Breathe app [50]. The processed vital parameters were also obtained from the Apple Health app, [31] which generates XML files containing information on processed parameters (e.g., heart rate, steps, sleep) and geolocation data when a workout is initiated.

Regarding Fitbit devices, data were downloaded from the Fitbit web platform [26], [33], or by the Pulse Watch mobile application [24].

Lastly, data recorded by Samsung [42] and Garmin [31] devices were downloaded directly via the manufacturer's application or the web portal. In [24], the Garmin device was connected to a computer via USB and then a.fit file was downloaded for subsequent processing.

Finally, energy expenditure estimates done with Apple watches and Garmin devices, were obtained directly from screenshots of the respective application [41].

Table 3 summarizes the extraction modalities reported across included studies, distinguishing between exported processed outcomes and availability of raw sensor data. In many consumer ecosystems, studies relied on processed outcomes (e.g., heart rate, steps, sleep summaries) rather than raw accelerometer or PPG signals; this should be explicitly documented when evidence is collected and interpreted.

4. Discussion

We focused on the main domains that could be useful for a forensic analysis and that are the most commonly extracted from smartwatches: movement, heart rate, and sleep.

Interestingly, all validation studies we found in the literature were in the health/clinical field, and none of them in the forensic one, mainly because smartwatches focus on the health/fitness of the user and because of the relatively new field of wearable forensics. As an example

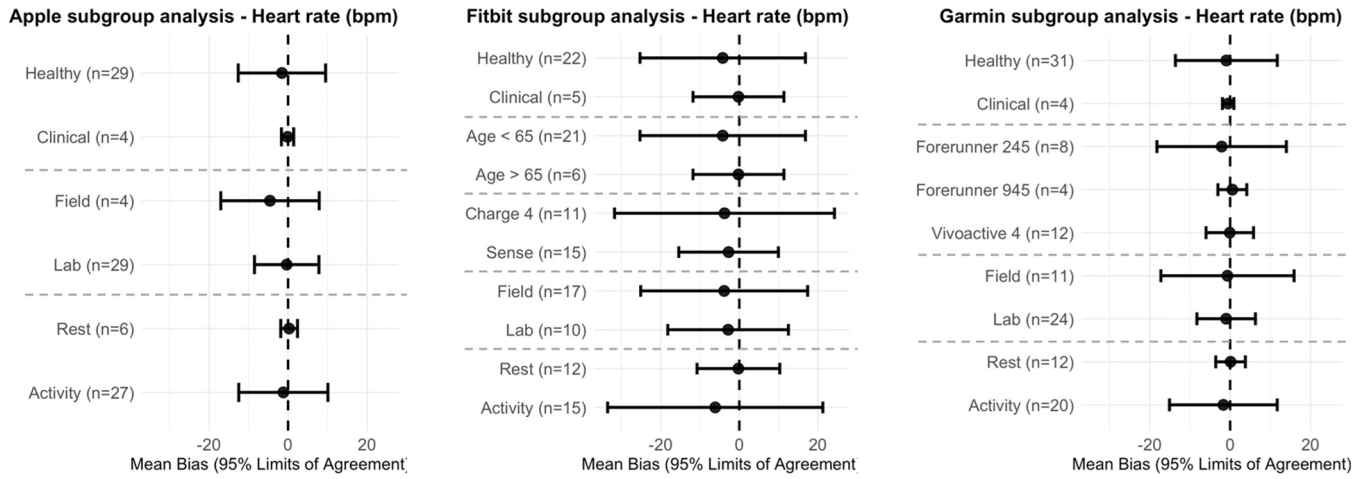

Fig. 7. Forest plots for the subgroup analyses of heart rate for Apple (left), Fitbit (middle), and Garmin (right) devices. N indicates the number of available comparisons between the tested device and the gold standard.

of a possible mismatch between health-oriented outcomes and outcomes that could be useful from a forensic perspective, Van Zandwijk et al. [12] found that iPhones detect using stairs (because this is related to a fitness goal), while they do not detect using an elevator, which is not interesting from a fitness perspective, but may be useful from a forensic point of view.

From our meta-analysis, we found that heart rate bias (average difference with respect to the actual value) obtained from Apple, Garmin, and Samsung was less than 1 beat per minute, with a high number of available comparisons, and relatively narrow limits of agreement. This suggests that these devices are promising for providing valid information about heart rate in forensic cases, such as reconstructing a suspect's (or victim) heart rate values during specific times. A confidence interval of $\pm$ 10 bpms on the value provided by the device was found. For Samsung watches, it should be considered that much fewer comparisons were available and therefore the corresponding result may be less robust. For Fitbit devices on the other hand, the more substantial mismatch (underestimation) of HR, along with broader limits of agreement, reflects a lower accuracy, which should be taken into account when using those values.

Considering previous reviews, accurate values of heart rate from Apple and Garmin are in line with a previous review by Fuller et al. [6] and the underestimation of heart rate from Fitbit devices is in line with both the review from Fuller et al. [6] and the meta-analysis by Chevance et al. [7], who found a very similar value of average underestimation (-2.99 bpm vs the current -3.44 bpm).

Only two studies were present for heart rate variability, not allowing a proper meta-analysis. Further studies are therefore warranted to evaluate the validity of such an outcome.

An important limitation for PPG-derived outcomes is the limited reporting (and limited diversity) of skin tone across validation cohorts. Because optical signal quality can vary across skin tones, the generalizability of reported PPG accuracy to diverse forensic cases may be reduced when Fitzpatrick skin type is not reported or when cohorts are not diverse enough regarding skin tone. Therefore, when interpreting smartwatch heart rate or HRV in casework, analysts should consider user characteristics as potential sources of additional uncertainty.

Considering the other outcomes, discrepancies were observed in the measurements of energy expenditure, number of steps, distance, and sleep. These discrepancies should be taken into account when using these outcomes.

Energy expenditure was underestimated by Apple Watch and overestimated by Fitbit (no info was obtained from the other devices), with substantially large (with respect to the energy expenditure of common activities) confidence intervals. This is in line with a previous review [6], which found that the mean percentage error (in absolute value) of energy expenditure estimates was over 3% for more than 90% of the comparisons in laboratory conditions and was over 10% (in absolute value) for more than 80% of the comparisons in free-living conditions.

Considering steps, many articles were discarded because they did not use proper gold standards (e.g., they used single research grade inertial measurement units, IMUs, as a reference). Garmin showed the best performance, with an average difference of 5 steps with respect to the reference. For future research, although a hand tally counter or video can be difficult to use for validation in free-living, new technologies such as body-worn cameras [60] and multiple sensor systems [61] are available for future studies. Valid frameworks of technical validation of real-world walking measures are also available [62].

Distance was only evaluated in two studies for different brands, not allowing a proper meta-analysis of the results.

Regarding sleep, Apple watches overestimated total sleep time, and Fitbit underestimated total sleep time with a bias of 23.5 min and of -4.9 min, respectively. Confidence intervals were over two hours, showing discrepancies with the reference values that should be considered.

As a limitation of the current review, it was not possible to use the chosen meta-analysis framework to analyze the discrimination between sleep and wakefulness. However, we found that sleep was better detected than wakefulness, with the former (sensitivity) ranging across devices from 89% to 98% and the latter (specificity) ranging from 13% to 78%.

A possible explanation for the different performance in sleep and wakefulness is due to the fact that most algorithms identify sleep based on accelerometers and on a lack of (substantial) movement at the level of the wrist, where they are usually worn. This explains why they have generally good performance in detecting sleep, which generally entails a lack of wrist movements, whereas they may mistake quiet wakefulness with no wrist movements for sleep. The limited number of studies and the wide limits of agreement further emphasize the need for caution in using these metrics.

Activity recognition was not considered in any of the studies using smartwatches, emphasizing the need for future work in this area. However, opening the field to digital devices such as smartphones, a study using an iPhone [12] found the accuracy in identifying floors (flights of stairs) to be between 70% and 80%, with minor changes depending on speed and carrying location.

Overall, several patterns observed here are consistent with prior literature findings, including that heart rate is among the most accurate smartwatch-derived outcomes, whereas energy expenditure consistently shows poorer agreement with criterion measures. Agreement for other

outcomes (e.g., steps, distance, sleep) is more heterogeneous across devices and study settings, and performance often worsens in free-living compared with laboratory protocols. Our analysis extends this literature by focusing on recent device generations (≥2019), providing brand-stratified pooled agreement estimates where possible, and by summarizing how data were extracted (often incompletely reported), which is particularly relevant for reproducibility and forensic analyses.

Regarding data extraction, the methodology to extract outcomes from fitness trackers and smartwatches was not always reported, and it varied depending on the study, the device, the manufacturer, and the associated applications. Still, the download from the manufacturer's app or web portal was the most common. In case of need to extract those outcomes for a forensic case, further difficulties may arise (e.g., blocked device), and both open access and commercial solutions exist to download data for forensics purposes [63], [64], [65]. The current review did not focus on this aspect, which should be studied in future research work.

Considering all devices, it should be considered that, although validation occurred in a variety of activities, few studies have performed it in real-world settings, which would be more similar to the ones of interest for forensic cases. Indeed, most included validation studies were conducted in laboratory settings, which improves experimental control but may underestimate the uncertainty encountered in real-world use. In free-living conditions, smartwatch signals are affected by a broader and less predictable range of movements (e.g., abrupt wrist motion and impacts), which can alter mobility estimation and increase motion artifacts in optical sensing. Importantly, human mobility itself is known to differ between laboratory and real-world contexts [66], [67]: walking in the lab tends to be more regular and constrained, whereas free-living walking is more variable in speed, turning, surface, and posture, potentially changing the relationship between wrist motion and the target quantity (e.g., steps, distance, and energy expenditure).

It should also be noted that for some outcomes, different computations (for heart rate variability) or different measurements (energy expenditure) were used, limiting comparisons and generalization of the results.

The constant evolution of the models and related software, the general lack of transparency from manufacturers about their proprietary (black box) algorithms, and the different available modalities of data extraction may be a limiting factor for the proper use of these technologies for forensic analyses.

Another topic that was not considered by the current review, but that could be useful for forensic analysis, is the possibility to alter/fabricate on purpose the outcomes provided by the smartwatch. Connected to this topic, one article of the review [37] reported that during a card game (sedentary activity), the smartwatch provided several (wrongly detected) steps, which suggests that step identification could be triggered on purpose or when doing other activities.

In this review, we focused on some of the most common metrics provided by a smartwatch. Future studies could focus on additional ones, such as position and path, and more recent ones, not common to all devices, such as breathing rate, oxygen saturation, maximal oxygen uptake, atrial fibrillation, electrodermal activity, blood pressure, and possibly further interesting features, such as fall detection. Also, similar analyses could be performed among other types of devices of the ever-growing wearable field, such as smartphones, smart glasses, rings, and virtual reality headsets.

A subset of the selected studies also reported mean percentage errors. Reporting this metric in future studies, together with bias and limits of agreement, may help quantify better the accuracy of the outcomes provided by smartwatches.

As a limitation of this review, we only focused on recent smartwatches and studies in order to obtain an updated picture of the validity of smartwatches that people are currently wearing and to focus on the latest versions of the manufacturer's algorithms. For future similar studies, it could be useful to find a rigorous way to incorporate evidence from previous studies and reviews focusing on previous models, keeping in mind that the algorithms were most likely updated in the meantime.

Moreover, we excluded participants younger than 14 years to align the review scope with common age thresholds for criminal responsibility in several jurisdictions. Nevertheless, the validity and interpretation of smartwatch-derived outcomes in individuals < 14 years is an important topic for future work, with potential legal relevance both when the young individual is a suspect/defendant and when he/she is a victim.

Finally, none of the studies had the lowest quality, but none of the studies had the highest quality either. To improve future validation and meta-analysis studies, we report below some suggestions (based on the articles we studied) to provide more useful, generalizable, and comparable information:

- Report complete information on the outcome values both of the reference and of the tested device (i.e., mean and SD of the outcome from the device and from the gold standard in the population considered)
- Report details on how the data was extracted and which data it was possible to extract (e.g. raw or processed data)
- Report in detail the characteristics of the protocol.
- Update software and firmware to the latest version and report the details of the used software and firmware of the devices, deactivating, if possible, possible updates during the study [31].
- Report at least bias, limits of agreement, and mean percentage errors (mean and SD). Bias and limits of agreement should be reported as values, not only in the graphical format of Bland-Altman plots.
- Report the number of subjects and the number of data points on which the comparison is performed.
- Report what difference was considered: ("tested device – gold standard value") or ("gold standard value – tested device"). In the majority of the studies of this review the first option, which we suggest, was used.
- Report both kcal and MET for energy expenditure when possible
- Report details on the wearing position
- Record and report if possible information that can impact specific outcomes (e.g. Fitzpatrick skin color scale and degree of forearm hair density for PPG; height for steps and distance). Follow available guidelines when possible [68]
- If assessing the performance of wake-sleep classification, follow a standardized framework including discrepancy analysis, Bland–Altman plots, and epoch-by-epoch analysis [69]
- Compare results with recognized gold standard measurements

5. Conclusions

We reviewed and meta-analysed the most recent smartwatch devices and their performance for activity classification, steps, distance walked, sleep, heart rate, and heart rate variability.

In our review, we found that not all brands of smartwatches and not all outcomes were equally studied, with varying results in terms of accuracy. Heart rate resulted the most studied (and most accurate) measure. Most studies validated smartwatches in healthy populations and performed validation in laboratory settings. For this reason, accuracy estimates may represent an upper bound, as real-world use can introduce additional movement-related noise and uncertainty in mobility patterns), which should be considered when interpreting smartwatch outputs in forensic contexts.

The presented results can provide useful insights into the performance of commercial devices for the forensic and clinical and wellness (health, fitness, sport) fields.

Based on the results of this review, we believe that future digital forensic (and health) research should focus on further validation studies, increasing sample size and performing real-world validation, on enhancing (and standardizing) methodologies for validating commercial devices, and advocating for transparency and standardization among

manufacturers, in order to support a valid forensic (and clinical) use of such technology.

CRediT authorship contribution statement

Lorenzo Chiari: Supervision, Conceptualization. **Alessandro Silvani:** Writing – review & editing, Methodology. **D'Ascanio Ilaria:** Methodology, Data curation. **Luca Palmerini:** Writing – review & editing, Supervision, Conceptualization. **Alberto Camon:** Supervision, Funding acquisition, Conceptualization. **Jose Albites-Sanabria:** Methodology, Data curation. **BARTOLI LAURA:** Writing – original draft, Resources, Project administration, Investigation, Funding acquisition. **Marcello Sicbaldi:** Writing – review & editing, Writing – original draft, Resources, Investigation, Data curation.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests. L. P. and L.C. are co-founders and shareholders of mHealth Technologies srl.

Acknowledgement

This work was supported by the PRIN project "New technologies, biometric data and criminal proceedings" MUR 2022CWNCH8 - CUP J53D23005360006, Call 2022.

Multimedia Appendix 1

Framework details, sensitivity analysis, and analysis without outliers (Multimedia Appendix 1.docx).

Multimedia Appendix 2

Data extracted from the articles, divided by brand (Multimedia Appendix 2.xlsx).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.forsciint.2026.112901.

References

[1] N.A. Almubairik, F.A. Khan, R.M. Mohammad, M. Alshahrani, WristSense framework: exploring the forensic potential of wrist-wear devices through case studies, For. Sci. Int. Digital Investig. 52 (Mar. 2025) 301862, https://doi.org/10.1016/J.FSIDI.2025.301862.
[2] A. MacDermott, S. Lea, F. Iqbal, I. Idowu, B. Shah, Forensic analysis of wearable devices: Fitbit, Garmin and HETP Watches (Jun), 2019 10th IFIP Int. Conf. N. Technol. Mobil. Secur. NTMS 2019 Proc. Workshop (2019), https://doi.org/10.1109/NTMS.2019.8763834.
[3] M. Schlusche, K. Yen, S. Knödler, K. Feld, Digitale spuren in einem mordprozess (Sep), Rechtsmed. 2025 356 35 (6) (2025) 446–450, https://doi.org/10.1007/s00194-025-00796-y.
[4] "Apple health data used in murder trial." Accessed: Feb. 04, 2026. [Online]. Available: ⟨https://www.bbc.com/news/technology-42663297⟩.
[5] F. Germini, et al., Accuracy and acceptability of wrist-wearable activity-tracking devices: systematic review of the literature, J. Med. Internet Res. 24 (1) (Jan. 2022), https://doi.org/10.2196/30791.
[6] D. Fuller, et al., Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review, JMIR Mhealth Uhealth 8 (9) (Sep. 2020), https://doi.org/10.2196/18694.
[7] G. Chevance, et al., Accuracy and precision of energy expenditure, heart rate, and steps measured by combined-sensing fitbits against reference measures: systematic review and meta-analysis, JMIR Mhealth Uhealth 10 (4) (Apr. 2022) e35626, https://doi.org/10.2196/35626.
[8] K.R. Evenson, C.L. Spade, Review of validity and reliability of garmin activity trackers, J. Meas. Phys. Behav. 3 (2) (2020) 170–185, https://doi.org/10.1123/JMPB.2019-0035.
[9] J.K. Lu, M. Sijm, G.E. Janssens, J. Goh, A.B. Maier, Remote monitoring technologies for measuring cardiovascular functions in community-dwelling adults: a systematic review, Geroscience 45 (5) (Oct. 2023) 2939–2950, https://doi.org/10.1007/S11357-023-00815-4.
[10] T. Lee, et al., Accuracy of 11 wearable, nearable, and airable consumer sleep trackers: prospective multicenter validation study, JMIR Mhealth Uhealth 11 (2023), https://doi.org/10.2196/50983.
[11] A.M. Schyvens, et al., Accuracy of Fitbit Charge 4, Garmin Vivosmart 4, and WHOOP Versus Polysomnography: Systematic Review, JMIR Mhealth Uhealth 12 (1) (Jan. 2024) e52192, https://doi.org/10.2196/52192.
[12] J.P. van Zandwijk, K. Lensen, A. Boztas, Have you been upstairs? On the accuracy of registrations of ascended and descended floors in iPhones, Forensic Science International Digital Investigation 47 (Dec. 2023) 301660, https://doi.org/10.1016/J.FSIDI.2023.301660.
[13] C. Werner, et al., Validity and reliability of the Apple Health app on iPhone for measuring gait parameters in children, adults, and seniors, *2023 13:1*, Sci. Rep. 13 (1) (2023) 1–11, https://doi.org/10.1038/s41598-023-32550-3.
[14] Apple Inc, Measuring Walking Quality Through iPhone Mobility Metrics (2021).
[15] M. de Zambotti, et al., Rigorous performance evaluation (previously, 'validation') for informed use of new technologies for sleep health measurement, Sleep. Health 8 (3) (Jun. 2022) 263–269, https://doi.org/10.1016/j.sleh.2022.02.006.
[16] J.F. Thayer, F. Åhs, M. Fredrikson, J.J. Sollers, T.D. Wager, A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health, Neurosci. Biobehav. Rev. 36 (2) (Feb. 2012) 747–756, https://doi.org/10.1016/J.NEUBIOREV.2011.11.009.
[17] R. Gilgen-Ammann, T. Schweizer, T. Wyss, Accuracy of the multisensory wristwatch polar vantage's estimation of energy expenditure in various activities: instrument validation study, JMIR Mhealth Uhealth 7 (10) (2019), https://doi.org/10.2196/14534.
[18] L.B. Mokkink, et al., The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study, Qual. Life Res. 19 (4) (May 2010) 539–549, https://doi.org/10.1007/S11136-010-9606-8.
[19] E. Tipton, J. Shuster, A framework for the meta-analysis of Bland-Altman studies based on a limits of agreement approach, Stat. Med. 36 (23) (Oct. 2017) 3621–3635, https://doi.org/10.1002/SIM.7352.
[20] M.J. Page, et al., The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, BMJ 372 (Mar. 2021), https://doi.org/10.1136/BMJ.N71.
[21] M. Budig, M. Keiner, R. Stoohs, M. Hoffmeister, V. Höltke, Heart rate and distance measurement of two multisport activity trackers and a cellphone app in different sports: a cross-sectional validation and comparison field study (Jan), Sensors 22 (1) (2022), https://doi.org/10.3390/S22010180.
[22] M. Budig, R. Stoohs, M. Keiner, Validity of two consumer multisport activity tracker and one accelerometer against polysomnography for measuring sleep parameters and vital data in a laboratory setting in sleep patients, Sensors 22 (23) (Dec. 2022), https://doi.org/10.3390/S22239540.
[23] M. Ceugniez, H. Devanne, E. Hermand, Reliability and accuracy of the fitbit charge 4 photoplethysmography heart rate sensor in ecological conditions: validation study, JMIR Mhealth Uhealth 13 (2025) e54871, https://doi.org/10.2196/54871.
[24] G. Colonna, et al., Measuring heart rate accurately in patients with parkinson disease during intense exercise: usability study of fitbit charge 4, JMIR Biomed. Eng. 8 (1) (Dec. 2023) e51515, https://doi.org/10.2196/51515.
[25] G. Cosoli, L. Antognoli, V. Veroli, L. Scalise, Accuracy and Precision of Wearable Devices for Real-Time Monitoring of Swimming Athletes, *2022, Vol. 22, Page 4726*, Sensors 22 (13) (2022) 4726, https://doi.org/10.3390/S22134726.
[26] X. Dong, S. Yang, Y. Guo, P. Lv, M. Wang, Y. Li, Validation of Fitbit Charge 4 for assessing sleep in Chinese patients with chronic insomnia: A comparison against polysomnography and actigraphy (October), PLoS One 17 (10) (Oct. 2022), https://doi.org/10.1371/JOURNAL.PONE.0275287.
[27] O.M. Giggins, et al., Accuracy of Wrist-Worn Photoplethysmography Devices at Measuring Heart Rate in the Laboratory and during Free-Living Activities, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2021, 2021, pp. 6970–6973, https://doi.org/10.1109/EMBC46164.2021.9629522.
[28] M.A. Grandner, et al., Performance of a multisensor smart ring to evaluate sleep: in-lab and home-based evaluation of generalized and personalized algorithms, Sleep 46 (1) (Jan. 2023), https://doi.org/10.1093/SLEEP/ZSAC152.
[29] G. Hajj-Boutros, M.A. Landry-Duval, A.S. Comtois, G. Gouspillou, A.D. Karelis, Wrist-worn devices for the measurement of heart rate and energy expenditure: A validation study for the Apple Watch 6, Polar Vantage V and Fitbit Sense, Eur. J. Sport Sci. 23 (2) (Feb. 2023) 165–177, https://doi.org/10.1080/17461391.2021.2023656.
[30] P. Helmer, et al., Accuracy and Systematic Biases of Heart Rate Measurements by Consumer-Grade Fitness Trackers in Postoperative Patients: Prospective Clinical Trial, J. Med. Internet Res. 24 (12) (Dec. 2022), https://doi.org/10.2196/42359.
[31] P. Helmer, et al., Reliability of continuous vital sign monitoring in post-operative patients employing consumer-grade fitness trackers: a randomised pilot trial, Digit. Health 10 (Jan. 2024), https://doi.org/10.1177/20552076241254026.
[32] E. Hermand, C. Coll, J.P. Richalet, F.J. Lhuissier, Accuracy and reliability of pulse O 2Saturation measured by a wrist-worn oximeter, Int. J. Sports Med. 42 (14) (Dec. 2021) 1268–1273, https://doi.org/10.1055/A-1337-2790.
[33] F. Hermans, et al., Validity of a consumer-based wearable to measure clinical parameters in patients with chronic obstructive pulmonary disease and healthy controls: observational study, JMIR Mhealth Uhealth 12 (Nov. 2024) e56027, https://doi.org/10.2196/56027.

[34] W.Te Ho, Y.J. Yang, T.C. Li, Accuracy of wrist-worn wearable devices for determining exercise intensity, Digit. Health 8 (Sep. 2022), https://doi.org/10.1177/20552076221124393/ASSET/IMAGES/LARGE/10.1177_20552076221124393-FIG4.JPEG.
[35] M. Jachymek, M.T. Jachymek, R.M. Kiedrowicz, J. Kaźmierczak, E. Płońska-Gościniak, M. Peregud-Pogorzelska, Wristbands in home-based rehabilitation—Validation of heart rate measurement, Sensors 22 (1) (Jan. 2022), https://doi.org/10.3390/S22010060.
[36] A. Jamieson, S. Jones, N. Chaturvedi, A.D. Hughes, M. Orini, Accuracy of smartwatches for the remote assessment of exercise capacity, *2024 14:1*, Sci. Rep. 14 (1) (Oct. 2024) 1–12, https://doi.org/10.1038/s41598-024-74140-x.
[37] K. Kastelic, M. Dobnik, S. Löfler, C. Hofer, N. Šarabon, Validity, Reliability and Sensitivity to Change of Three Consumer-Grade Activity Trackers in Controlled and Free-Living Conditions among Older Adults, Sens. (Basel) 21 (18) (Sep. 2021) 6245, https://doi.org/10.3390/S21186245.
[38] A.A. Khushhal, A.A. Mohamed, M.E. Elsayed, Accuracy of Apple Watch to Measure Cardiovascular Indices in Patients with Chronic Diseases: A Cross Sectional Study, J. Multidiscip. Health 17 (2024) 1053–1063, https://doi.org/10.2147/JMDH.S449071.
[39] C. Kim, S.H. Kim, M.R. Suh, Accuracy and Validity of Commercial Smart Bands for Heart Rate Measurements During Cardiopulmonary Exercise Test, Ann. Rehabil. Med. 46 (4) (Aug. 2022) 209–218, https://doi.org/10.5535/ARM.22050.
[40] C. Kim, J.H. Song, S.H. Kim, Validation of Wearable Digital Devices for Heart Rate Measurement During Exercise Test in Patients With Coronary Artery Disease, Ann. Rehabil. Med. 47 (4) (2023) 261, https://doi.org/10.5535/ARM.23019.
[41] S. Le, et al., Validity of three smartwatches in estimating energy expenditure during outdoor walking and running, Front. Physiol. 13 (Sep. 2022), https://doi.org/10.3389/FPHYS.2022.995575.
[42] T. Lee, et al., Accuracy of 11 Wearable, Nearable, and Airable Consumer Sleep Trackers: Prospective Multicenter Validation Study, JMIR Mhealth Uhealth 11 (2023), https://doi.org/10.2196/50983.
[43] Z. Li, W. Feng, L. Zhou, S. Gong, Accuracy of wrist-worn activity trackers for measuring steps in patients after major abdominal surgery: A validation study, Digit. Health 10 (Jan. 2024), https://doi.org/10.1177/20552076241297036.
[44] S.E. Lim, H.S. Kim, S.W. Lee, K.H. Bae, Y.H. Baek, Validation of Fitbit Inspire 2™ Against Polysomnography in Adults Considering Adaptation for Use, Nat. Sci. Sleep. 15 (Feb. 2023) 59–67, https://doi.org/10.2147/NSS.S391802.
[45] C.S. Lima, F.C. Bertocco, J.I.V. De Oliveira, T.M.F. De Souza, E.P. Da Silva, F.J. Von Zuben, Assessment of Samsung Galaxy Watch4 PPG-Based Heart Rate during Light-to-Vigorous Physical Activities, IEEE Sens. Lett. 8 (7) (Jul. 2024), https://doi.org/10.1109/LSENS.2024.3408089.
[46] C. Meza, et al., Accuracy of a Smartwatch to Assess Heart Rate Monitoring and Atrial Fibrillation in Stroke Patients, *2023, Vol. 23, Page 4632*, Sensors 23 (10) (May 2023) 4632, https://doi.org/10.3390/S23104632.
[47] D.J. Miller, C. Sargent, G.D. Roach, A Validation of Six Wearable Devices for Estimating Sleep, Heart Rate and Heart Rate Variability in Healthy Adults, Sensors 22 (16) (Aug. 2022) 6317, https://doi.org/10.3390/S22166317.
[48] M. Neudorfer, et al., Validity of Four Consumer-Grade Optical Heart Rate Sensors for Assessing Volume and Intensity Distribution of Physical Activity, Scand. J. Med. Sci. Sports 34 (11) (Nov. 2024), https://doi.org/10.1111/SMS.14756.
[49] M. Ogasawara, et al., Exploratory Validation of Sleep-Tracking Devices in Patients with Psychiatric Disorders, Nat. Sci. Sleep. 15 (2023) 301–312, https://doi.org/10.2147/NSS.S400944.
[50] B. O'Grady, R. Lambe, M. Baldwin, T. Acheson, C. Doherty, The validity of apple watch series 9 and ultra 2 for serial measurements of heart rate variability and resting heart rate, *2024, Vol. 24, Page 6220*, Sensors 24 (19) (Sep. 2024) 6220, https://doi.org/10.3390/S24196220.
[51] J.L. Ong, et al., Selecting a sleep tracker from EEG-based, iteratively improved, low-cost multisensor, and actigraphy-only devices, Sleep. Health 10 (1) (Feb. 2024) 9–23, https://doi.org/10.1016/j.sleh.2023.11.005.
[52] R. Robbins, et al., Accuracy of Three Commercial Wearable Devices for Sleep Tracking in Healthy Adults, Sensors 24 (20) (Oct. 2024), https://doi.org/10.3390/S24206532.
[53] L. Simón-Vicente, et al., Validation of ActiGraph and Fitbit in the assessment of energy expenditure in Huntington's disease, Gait Posture 109 (Mar. 2024) 89–94, https://doi.org/10.1016/j.gaitpost.2024.01.028.
[54] M.P. Støve, E.C.K. Hansen, Accuracy of the Apple Watch Series 6 and the Whoop Band 3.0 for assessing heart rate during resistance exercises, J. Sports Sci. 40 (23) (2022) 2639–2644, https://doi.org/10.1080/02640414.2023.2180160.
[55] J. Stutz, et al., Energy expenditure estimation during activities of daily living in middle-aged and older adults using an accelerometer integrated into a hearing aid, Front. Digit. Health 6 (Jun. 2024) 1400535, https://doi.org/10.3389/FDGTH.2024.1400535/BIBTEX.
[56] X. Sun, Z. Wang, X. Fu, C. Zhao, F. Wang, H. He, Validity of Apple Watch 6 and Polar A370 for monitoring energy expenditure while resting or performing light to vigorous physical activity, J. Sci. Med. Sport 26 (9) (Sep. 2023) 482–486, https://doi.org/10.1016/j.jsams.2023.07.005.
[57] F. Theurl, et al., Smartwatch-derived heart rate variability: a head-to-head comparison with the gold standard in cardiovascular disease, Eur. Heart J. Digit. Health 4 (3) (May 2023) 155–164, https://doi.org/10.1093/EHJDH/ZTAD022.
[58] J. Viciana, C. Casado-Robles, S. Guijarro-Romero, D. Mayorga-Vega, Are Wrist-Worn Activity Trackers and Mobile Applications Valid for Assessing Physical Activity in High School Students? Wearfit Study, J. Sports Sci. Med. 21 (3) (Sep. 2022) 356–375, https://doi.org/10.52082/JSSM.2022.356.
[59] A.R. Willoughby, et al., Performance of wearable sleep trackers during nocturnal sleep and periods of simulated real-world smartphone use, Sleep. Health 10 (3) (Jun. 2024) 356–368, https://doi.org/10.1016/J.SLEH.2024.02.007/ASSET/38FABC7A-1B16–456 C-B598-27A116E06C1B/MAIN.ASSETS/GR11.JPG.
[60] A.K. Bourke, E.A.F. Ihlen, R. Bergquist, P.B. Wik, B. Vereijken, J.L. Helbostad, A physical activity reference data-set recorded from older adults using body-worn inertial sensors and video technology-the ADAPT study data-set, Sensors 17 (3) (Mar. 2017), https://doi.org/10.3390/S17030559.
[61] F. Salis, et al., A multi-sensor wearable system for the assessment of diseased gait in real-world conditions, Front. Bioeng. Biotechnol. 11 (2023), https://doi.org/10.3389/fbioe.2023.1143248.
[62] C. Mazzà, et al., Technical validation of real-world monitoring of gait: A multicentric observational study, BMJ Open 11 (12) (2021), https://doi.org/10.1136/bmjopen-2021-050785.
[63] A. MacDermott, S. Lea, F. Iqbal, I. Idowu, B. Shah, Forensic analysis of wearable devices: Fitbit, Garmin and HETP Watches, 2019 10th IFIP Int. Conf. N. Technol. Mobil. Secur. NTMS 2019 Proc. Workshop (Jun. 2019), https://doi.org/10.1109/NTMS.2019.8763834.
[64] A. Almogbil, A. Alghofaili, C. Deane, T. Leschke, A. Almogbil, A. Alghofaili, Digital Forensic Analysis of Fitbit Wearable Technology: An Investigator's Guide. Proceedings - 2020 7th IEEE International Conference on Cyber Security and Cloud Computing and 2020 6th IEEE International Conference on Edge Computing and Scalable Cloud, CSCloud-EdgeCom 2020, Aug. 2020, pp. 44–49, https://doi.org/10.1109/CSCLOUD-EDGECOM49738.2020.00017.
[65] J. Williams, A. Macdermott, K. Stamp, F. Iqbal, Forensic Analysis of Fitbit Versa: Android vs iOS. Proceedings - 2021 IEEE Symposium on Security and Privacy Workshops, SPW 2021, May 2021, pp. 318–326, https://doi.org/10.1109/SPW53761.2021.00052.
[66] S. Del Din, A. Godfrey, B. Galna, S. Lord, L. Rochester, Free-living gait characteristics in ageing and Parkinson's disease: Impact of environment and ambulatory bout length, J. Neuroeng. Rehabil. 13 (1) (2016), https://doi.org/10.1186/s12984-016-0154-5.
[67] V.V. Shah, et al., Laboratory versus daily life gait characteristics in patients with multiple sclerosis, Parkinson's disease, and matched controls, J. Neuroeng. Rehabil. 17 (1) (Dec. 2020) 159, https://doi.org/10.1186/s12984-020-00781-4.
[68] B.W. Nelson, C.A. Low, N. Jacobson, P. Areán, J. Torous, N.B. Allen, Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research, NPJ Digit. Med. 3 (1) (Dec. 2020), https://doi.org/10.1038/S41746-020-0297-4.
[69] L. Menghini, N. Cellini, A. Goldstone, F.C. Baker, M. De Zambotti, A standardized framework for testing the performance of sleep-Tracking technology: step-by-step guidelines and open-source code, Sleep 44 (2) (Feb. 2021), https://doi.org/10.1093/SLEEP/ZSAA170.