OPEN ACCESS

Recommendations for determining the validity of consumer wearable heart rate devices: expert statement and checklist of the INTERLIVE Network

Jan M Mühlen ,[1] Julie Stang,[2] Esben Lykke Skovgaard,[3] Pedro B Judice ,[4,5] Pablo Molina-Garcia ,[6] William Johnston ,[7,8] Luís B Sardinha ,[9] Francisco B Ortega ,[6,10] Brian Caulfield ,[7,8] Wilhelm Bloch,[1] Sulin Cheng,[1,11] Ulf Ekelund ,[2] Jan Christian Brønd ,[3] Anders Grøntved,[3] Moritz Schumann [1,11]

► Additional material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bjsports-2020-103148).

For numbered affiliations see end of article.

Correspondence to
Dr Moritz Schumann, Institute of Cardiovascular Research and Sports Medicine, Department of Molecular and Cellular Sports Medicine, German Sport University Cologne, Koln 50858, Germany;
m.schumann@dshs-koeln.de

JMM, JS and ELS contributed equally.

Accepted 24 November 2020
Published Online First 4 January 2021

To cite: Mühlen JM, Stang J, Lykke Skovgaard E, *et al*. *Br J Sports Med* 2021;**55**:767–779.

ABSTRACT

Assessing vital signs such as heart rate (HR) by wearable devices in a lifestyle-related environment provides widespread opportunities for public health related research and applications. Commonly, consumer wearable devices assessing HR are based on photoplethysmography (PPG), where HR is determined by absorption and reflection of emitted light by the blood. However, methodological differences and shortcomings in the validation process hamper the comparability of the validity of various wearable devices assessing HR. Towards Intelligent Health and Well-Being: Network of Physical Activity Assessment (INTERLIVE) is a joint European initiative of six universities and one industrial partner. The consortium was founded in 2019 and strives towards developing best-practice recommendations for evaluating the validity of consumer wearables and smartphones. This expert statement presents a best-practice validation protocol for consumer wearables assessing HR by PPG. The recommendations were developed through the following multi-stage process: (1) a systematic literature review based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses, (2) an unstructured review of the wider literature pertaining to factors that may introduce bias during the validation of these devices and (3) evidence-informed expert opinions of the INTERLIVE Network. A total of 44 articles were deemed eligible and retrieved through our systematic literature review. Based on these studies, a wider literature review and our evidence-informed expert opinions, we propose a validation framework with standardised recommendations using six domains: considerations for the target population, criterion measure, index measure, testing conditions, data processing and the statistical analysis. As such, this paper presents recommendations to standardise the validity testing and reporting of PPG-based HR wearables used by consumers. Moreover, checklists are provided to guide the validation protocol development and reporting. This will ensure that manufacturers, consumers, healthcare providers and researchers use wearables safely and to its full potential.

INTRODUCTION

Heart rate (HR) is defined as the number of heart beats per minute (bpm) and can be determined from the time interval between two successive cardiac cycles initiated by action potentials in the sinoatrial node.[1] While resting HR is a key vital sign and a well-established predictor of all-cause and cardiovascular mortality in the general population,[2] other features of HR such as the response to exercise and HR variability (HRV) are indicators of general health status, including fitness as well as both physiological and mental stress.[3–5] Furthermore, HR assessment during exercise training is an important tool for monitoring training load in elite athletes and recreational exercisers.[6 7]

Traditionally, HR is derived from electrocardiography (ECG) recordings through either multiple-lead channels or simple chest-straps, consisting of two electrodes. Thus, HR assessment has traditionally been limited to medical conditions, laboratory testing or training monitoring and was not suitable for long-term assessment during daily living. However, recently a wealth of wearables that assess HR by photoplethysmography (PPG) have entered the consumer market. This allows not only for continuous fitness monitoring, but also facilitates screening for incident disease and continuous monitoring of disease progression and complications (eg, detection of atrial fibrillation and stroke prevention, coronary artery disease or sleep apnoea).[8–13]

PPG is an optical technique that is based on the absorption and reflection of emitted light by the blood, where the systolic variations in blood volume modulate the amount of transmitted or reflected light.[14] However, considerable differences in the validity of HR assessed by PPG-based devices are observed,[15] which are likely related to difficulties in mathematical peak detection and a higher sensitivity to motion artefacts.[16] This, in turn, may have severe consequences for long-term adherence to regular exercise,[17] but also for risk stratification if the device is used in a clinical setting.[18]

Unfortunately, the validation quality of wearables remains often unknown to the consumer due to non-transparent standards for testing and reporting. The validity assessment of consumer wearables is most optimally performed by independent institutions, but the number of new devices introduced by a continuously rising number of device manufacturers makes it almost impossible for scientific institutions to keep up with recent developments. Moreover, the discontinuation of a device or implementation of important changes to a device firmware/software might invalidate previous

work.[19] Therefore, it is important to develop a common framework for the optimal validity evaluation of consumer wearables measuring HR by PPG, to be used by both manufacturers and research institutions in order to provide quality assurance of available devices.

In 2018, the Consumer Technology Association published a preliminary framework for evaluating and reporting the validity for measuring HR with consumer wearables, including considerations for testing protocols but also individual characteristics, such as skin tone, body mass index (BMI), sex and age.[20] However, recommendations for long-term monitoring of HR during free-living conditions are lacking in these guidelines and the scientific evidence for the suggested guidelines has not been presented. In addition, in a recently published review article factors that may affect the accuracy of wrist-worn HR wearables were critically discussed and initial considerations for performing validity testing of these devices provided.[21] However, the published work mainly targets scientific evaluations of these devices and specific guidelines that allow for an immediate transfer into practice have not been presented.

Therefore, the present expert statement aims to expand on previously published work by proposing a set of guidelines targeting both manufacturers and scientific institutions, to ensure the rigorous and transparent validation and accuracy reporting of PPG-based consumer wearable HR devices, while at the same time being feasible to carry out. Furthermore, the statement aims to propose a best-practice framework of rigorousness in evaluating criterion validity and provide recommendations for future development of evaluating the validity of wearable HR monitors used by consumers. The work presented is based on a systematic literature search as well as an unstructured review of the wider literature pertaining to factors that may introduce bias during the validation of these devices and evidence informed expert opinions of the INTElligent Health and Well-being: NetwoRk of PhysicaL ActIVity AssEssment (INTERLIVE). As a result, we provide a comprehensive summary of variables that require consideration when developing evaluation protocols (online supplemental table 1) and suggest practical checklists for validation protocol designing (table 1) and transparent data reporting (table 2).

EXPERT STATEMENT PROCESS

The INTERLIVE Network

INTERLIVE is a joint initiative of the University of Lisbon (Portugal), the German Sport University (Germany), University of Southern Denmark (Denmark), Norwegian School of Sport Sciences (Norway), University College Dublin (Ireland), University of Granada (Spain) and Huawei Technologies Finland. The consortium was founded in 2019 and strives towards developing best-practice protocols for evaluating the validity of consumer wearables. Moreover, we are aiming to increase awareness of the advantages and limitations of different validation protocols and to introduce novel health-related metrics, fostering a wide-spread use of physical activity indicators. As one of the initial key aims of the group, the consortium aimed to develop best-practice validation protocols for consumer wearable HR monitoring (part A) and wearable and smartphone devices for step-counting (part B, presented in a separate publication).

Table 1 Checklist of items that need to be considered when planning validity protocols for consumer heart rate wearables

Target population	
BMI	O
Body height	O
Skin tone	O
Sex	O
Sample size calculation via pilot study	O
Criterion measure	
Chest strap or ECG	O
Placement according to manufacturer's instructions	O
Index device	
Placement according to manufacturer's instructions	O
Pretest preparations	
Standardised meal replacement	O
Control caffeine intake	O
Medical screening	O
Exclude participants with medication affecting cardiovascular function	O
Control for previous intense physical activity	O
Testing: laboratory conditions	
Minimum of 3 walking intensities	O
Minimum of 2 running intensities	O
Minimum of 3 biking intensities	O
Steady-state (2–5 min)	O
HR kinetics (transitions and recovery)	O
Validity level	
1. Graded ergometer test with a wide range of exercise intensities reported as % of HR_{max} (or VO_{2max}) including rest and recovery	O
2. Graded ergometer test with a wide range of exercise intensities in absolute values (ie, speed/incline, W/rpm) including rest and recovery	O
3. Graded ergometer test with a moderate range of exercise intensities as % of HR_{max} (or VO_{2max}) including rest and recovery	O
4. Graded ergometer test with a moderate range of exercise intensities reported in absolute values (ie, speed/incline, W/rpm) including rest and recovery	O
5. Graded ergometer test with a low range of exercise intensities reported as % of HR_{max} (or VO_{2max}) including rest and recovery	O
6. Graded ergometer test with a low range of exercise intensities reported in absolute values (ie, speed/incline, W/rpm) including rest and recovery	O
Testing: semifree-living (sport-specific) conditions	
Intermittent activities (ie, soccer, basketball, etc)	
1. Inherent environmental conditions (eg, standard playing field, etc)	O
2. Inherent no of players included	O
3. Inherent duration with a minimum of 15–20 min	O
Continuous activities (running, walking, biking, swimming)	
1. Minimum of three intensities (40 %, 60 %, 80% of HR_{max})	O
2. Inherent duration with a minimum of 2 min of each intensity	O
Activities with domestic behaviour (doing laundry, gardening, home construction)	
1. Minimum of 15–20 min	O
Testing: free-living conditions	
Subject's wear index and criterion device for a minimum of 24 hours	O
Exclude subject's not presenting HR data above 40% of HR_{max}	O
Exclude recordings missing more than 5% of the data (index or criterion device)	O
Processing	
Criterion measure processing	
1. Apply an automated method for filtering ectopic beats and motion artefacts	O
Index measure processing	
1. No post processing of the end-user data is allowed	O

Continued

Table 1 Continued

2. Resampling into a window of 5 s is allowed	O
Epochs for analysis/window size	
1. Sample criterion measure with same epoch as available with the index measure	O
2. Window size should be 5 s or shorter	O
Index and criterion synchronisation	
1. Automated method for synchronisation (cross correlation or similar)	O
Statistical analysis	
Standard Bland-Altman LoA analysis for steady-state conditions	O
Repeated measure LoA analysis for non-steady state conditions (multiple paired observations of HR epochs per individual)	O
Evaluate within-device precision by comparing the within-person variability in average HR over 5 s windows separately for steady-state activity of at least 2 min duration	O

BMI, body mass index; HR, heart rate; HR_{max}, maximal heart rate; LoA, limits of agreement; rpm, repetitions per minute; $VO2_{max}$, maximal oxygen uptake; W, Watts.

Expert validation protocol development

Expert validation process

An initial meeting was held in Cascais, Portugal on 15 November 2019. At this meeting, it was agreed that the optimal process for developing the best-practice validation protocol should begin with extracting key elements of the validation protocols previously used in the scientific literature. This information was then used as the foundation for discussions on the optimal and feasible protocol for conducting the validity assessment that describes the accuracy end-users can expect if the wearable is used in the designated or similar setting. The consortium formed two working groups: (1) HR monitoring (JMM, ELS, JS, SC, WB, JCB, UE, AG and MS), (2) step-counting (WJ, PMG, PBJ, BC, FBO and LBS). The working groups subsequently defined multiple systematic literature search strategies, prior to sharing them with the wider consortium. A second consortium meeting was held virtually on 10 March 2020 to finalise the search strategies, including the selection of the minimum a priori required criterion measure(s). Thereafter, the systematic search was performed and a framework was developed for extracting data of the validation process, including data on target population, criterion and index device, testing conditions, data processing and statistical analysis. In parallel, an unstructured review of the wider literature was conducted to include valid studies on factors that may affect the accuracy on consumer wearables not identified by our defined search strategies. Following that, the data extraction was performed and multiple workgroup meetings were held to discuss each aspect of the validation protocols used in the individual studies. Based on the data synthesised during the systematic literature review, the a priori knowledge relating to research grade device validation[22–25] and the evidence informed expert opinion of the INTERLIVE members, a set of key domains for the best-practice recommendations were proposed. The synthesised data were then reviewed with respect to these domains, and expert validation protocols for wearable HR monitors (part A) and wearable and smartphone step-counters (part B) and were iteratively developed by the working groups and subsequently shared with the entire consortium. At a virtual meeting held on 17 June 2020, the revised drafts were discussed and the two protocols were aligned to ensure harmonisation of the statements. The revised drafts were then edited for consistency and reviewed by the wider consortium prior to circulation for final approval.

Table 2 Minimum required reporting sheet for standardized and transparent data sharing

	Description	Reporting
Target population		
Sampling method	Random, convenient, and so on	
Distribution of sex	♂=n/♀=n	
Body height	Mean±SD and range (cm)	
BMI	Mean±SD and range (kg/m²)	
Skin tone	Distribution of Fitzpatrick scale	
Sample size	Number of subjects	
Criterion measure		
Chest strap or ECG (RR intervals)	Model and brand; chest strap: agreement with respect to bpm	
Placement	Manufacturer's instructions and actual placement	
Index device		
Placement	Manufacturer's instructions and actual placement	
Pre-test preparation		
Standardised meal replacement	Type of replacement and duration of control (hours prior to testing)	
Caffeine intake	Duration of control (hours prior to testing)	
Medical screening	Type of medical screening	
Exclusion of participants	Exclusions due to specific medication affecting cardiovascular function	
Intense physical activity	Duration of control (hours prior to testing)	
Testing protocol		
Type of protocol	Laboratory, semi-free living/sport-specific, free-living	
Contextual factors	Indoors, outdoors	
Type of activity	Cycling, treadmill walking/running, swimming, other sports/activities	
Duration	Minutes, hours	
Exercise intensity	Relative to aerobic capacity (%HR_{max}, VO_{2max}, RM) Or Absolute values (ie, speed/ incline, W/ rpm)	
Steady-state (2–5 min)	Mean HR (ie, 5–30 s intervals)	
HR kinetics (transitions and recovery)	ΔHR	
Processing		
Criterion measure processing	Method used for error correction and data smoothing	
Index measure processing	Method used for resampling (is used)	
Epochs for analysis/window size	In seconds	
Index and criterion synchronisation	Method used (cross-correlations or similar methods)	
Statistical analysis (report separately for each exercise intensity and/or activity)		
N of paired observations	Paired HR (amount)	
Data availability	Data availability (%)	
Index device, mean HR	Mean±SD	
Criterion device, mean HR	Mean±SD	
Mean difference	Mean±SD and SE	
Mean absolute error	Bpm	
MAPE	%	
Standard LoA	Mean difference or mean relative difference and LoA including 95% CIs (separately for each steady-state intensity and/or activity)	
Repeated measure LoA analysis	Mean difference or mean relative difference and LoA including 95% CIs (separately for each non-steady-state activities)	

Continued

Table 2 Continued

	Description	Reporting
Within–device precision for steady-state activities	The 95% prediction interval and ICC (report separately for each steady-state intensity and/or activity)	
Other		
Deviations in the validation process		

BMI, body mass index; bpm, beats per minute; ECG, electrocardiogram; HR, heart rate; HRmax, maximal heart rate; ICC, intra class correlation coefficient; LoA, limits of agreement; MAPE, mean absolute percentage error; RM, repetition maximum; rpm, repetitions per minute; SE, standard error; VO2max, maximal oxygen uptake; W, Watt.

Systematic review process

The primary aim of our initial systematic literature review was to determine which methods and protocols are currently used in the scientific literature to validate HR with consumer-based wearables. Importantly, we did not aim to review the results from studies examining the validity of wearable consumer devices to assess HR. The search was conducted with respect to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) and registered with the international database of prospectively registered systematic reviews in health and social care (PROSPERO ID: CRD42020177667). Three-domain search terms were used to identify journal articles published in the electronic databases PubMed, Embase and Web of Science. More specifically, these search terms were defined as the control device, the outcome as well as the study design (online supplemental table 2).

Only English language publications in human populations with no restriction to publication year were included. Relevant articles had to be published prior to 18 March 2020. The inclusion criteria were defined as Population, Intervention, Comparison, Outcome and Study design .[26] No restrictions were made with regards to population (ie, healthy, patients, children, etc) and interventions (ie, protocols used). Protocols were classified as (1) laboratory settings (ie, well-controlled conditions, including isolated tasks such as walking, running or cycling on a treadmill or a stationary cycle ergometer), (2) semifree-living settings (semi-controlled conditions, including 'simulated' activities of daily living for the purpose of replicating 'free-living' conditions) or (3) free-living settings (long-term monitoring of daily living without restrictions of the completed tasks). As a comparison, a criterion measure using a gold standard (ie, assessment of the time elapsed between two successive R-waves [RR intervals) of the signal of sequence of the Q, R and S complex [QRS]) was required. Furthermore, only studies that assessed HR by a PPG-based consumer wearable as the primary outcome measure were included. However, no restrictions applied to the human-wearable interface of the index devices (eg, light wavelength or measurement site). The detailed search string can be found in online supplemental table 2.

Screening and data extraction were performed independently by three members of the consortium, using Covidence software (Veritas Health Innovation). The search process entailed saving the online search, removing duplicates as well as consequently screening titles, abstracts and eligible full texts. A minimum of two identical votes was required for eligibility judgement. In case of a lack of consensus, the third member of the team was consulted. Data extraction was performed according to specific criteria that are outlined in detail in online supplemental tables 3–6.

CURRENT STATE OF KNOWLEDGE

The presented current state of knowledge is based on the studies that were identified by the systematic literature search as well as supplemental technological studies and our evidence informed expert opinions. Our systematic review led to a total of 1894 hits. Automatically removing duplicates and ineligible records led to 108 full texts for further assessment. Overall, 66 studies were primarily excluded for methodological reasons (ie, in terms of outcome, study design and comparator). Finally, a total of 43 articles were deemed eligible and retrieved. Additionally, one study was manually added by screening other resources, leading to a total number of 44 studies remaining for data extraction. The PRISMA flow chart of the systematic review process and the reasons for exclusions are presented in the supplements (online supplemental figure 1). The following section provides a short summary on the key considerations that appear to be important when testing the validity of PPG-based consumer wearables. Data gathered are presented in six key domains (figure 1), that were deemed relevant for validity testing (target population, criterion measure, device placement, testing conditions, data processing and statistical analysis). The most important aspects of validation protocols are also summarised in the supplements, table 1. The consortium acknowledges that the presented list of domains reflects the current state of knowledge but may not be considered exhaustive.

Target population

Selecting the target population for the validity assessment appears to be a key factor that determines the significance of the findings obtained. Although PPG-based wearables could theoretically be validated in numerous populations that differ considerably in demographics, ethnicity, anthropometrics and activity level, we advise that the evaluation reflects the device performance in the hands of the intended user. However, even by assessing the validity in a sample that is homogeneous in one domain (eg, recreationally active young men), it is likely that other domains may not be controlled for simultaneously (eg, skin tone). Therefore, we suggest that the target populations generally reflect a heterogeneous sample, allowing for possible subgroup analysis. Homogeneous samples, on the other hand, may be included if the intention is to test the validity of the wearable for a very specific group (eg, athletes of a specific sport).

In addition to the aforementioned considerations, other factors may require attention. For example, the pathology of some heart-related diseases may affect the outline of the QRS complex and potentially provide poor identification of the R wave.[27] However, the number of heart-related conditions and their potential implications on the QRS complex impose numerous challenges with the validity assessment. These challenges of including patients with heart-related conditions must be appropriately addressed to ensure the accuracy of the HR measurements. In this context, assessment of atrial fibrillation has been targeted by some wearable devices[28] and it seems possible that the future will bring more devices that can address specific heart-related conditions that can be detected with a high degree of confidence.[29] In addition, abnormalities in blood pressure may affect the PPG signal.[30] Consequently, in many studies included in our review, participants with known systolic or diastolic blood pressure abnormalities were excluded[31 32] or at least reported.[33–38]

Other considerations concern the use of medication and dietary supplements that may affect HR recordings and should be considered when designing validation protocols. Interestingly,

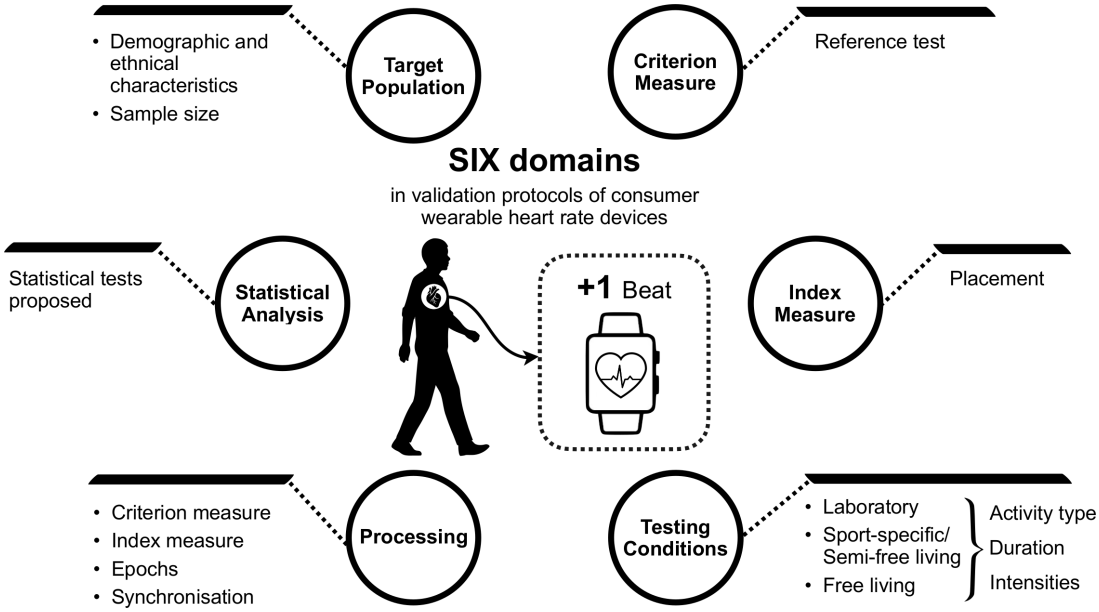


Figure 1 The six domains identified as important factors to be considered during validity testing of wearable devices assessing heart rate by photoplethysmography.

none of the identified studies assessed the accuracy of wearable HR monitors in patients with cardiovascular conditions. These patients present a variety of potential challenges to monitor's accuracy, including hypertension, peripheral arterial disease, venous insufficiency, obesity, atrial fibrillation and use of medications that affect HR, vascular tone and volume status (eg, beta-blockers, ACE inhibitors, calcium channel blockers, and diuretics).[33] Furthermore, factors such as inked and damaged skin (eg, tattoos, scars etc.) may potentially affect the PPG signal and, thus, participants exhibiting either of these were excluded in few studies identified by our systematic review.[33 39–41] While we also recommend that exclusion will likely help to overcome potential errors originating from these factors, in this statement we have focused on the validity assessment of HR measurements in the general population. Principally, healthy samples are recommended for the general device validation. However, if a device is specifically designed for a special population, this needs to be reflected in the target population.

The following factors require consideration when designing an appropriate validation protocol.

Sample size considerations

The sample size should be defined a priori. If an a priori specified level of 'in agreement' (ie, the difference of paired measurements falls within a specified interval) is considered, the sample size should be calculated based on an expected mean absolute difference, the expected standard deviation (SD) of the differences, and a predefined clinical maximum allowed difference needed to obtain a power of 80% or 90% to assess agreement between two methods of measurement with a sufficient precision.[42] It is advised to conduct a pilot study to obtain the mean and SD of differences between the wearable consumer device and the criterion measure to make these prior sample size calculations. If no a priori specified level of 'in agreement' is considered, for homogeneous samples we recommend a minimum of 45 participants as a rule of thumb.[43] This number is also in line with the average number of participants included in the studies identified by our systematic review. In any case, the variability of relevant participant characteristics in the sample should be considered and for heterogeneous groups, a larger sample size might be necessary.

Age

Ageing has previously been associated with increases in arterial stiffness, resulting in changes in the propagation of the pulse to the periphery, thereby affecting pulse timing and shape characteristics.[44] However, only three studies identified by our systematic literature review performed a statistical analysis for age and device error (total range 21–73 years), but did not find age to affect the error in the prediction of HR measurements.[41 45 46] In fact, this finding was confirmed by a very recent study validating the wearable fitness trackers Xiaomi Mi Band 2 and Garmin Vivosmart HR+.[47] Also, in this study, similar mean percentage errors for young (20–26 years) participants and seniors (>65 years) were observed. Thus, deteriorating effects on vascular function with increasing age may not be reflected in HR assessed by PPG but more research is needed to clearly assess these effects. However, if the validity of a wearable device is not needed for one specific target group (eg, children) we suggest testing in a heterogeneous sample.

Sex

Sex is associated with cardiovascular function, affecting resting HR and arterial blood pressure.[48] Consequently, sex differences might also be reflected in PPG-based HR due to possible differences in device positioning and skin characteristics.[45] For example, differences seem to exist in the thickness and echo intensity of skin between males and females.[49] However, three studies identified by our systematic literature search did not find HR validity assessed via PPG to be affected by sex,[32 39 41] while others[45 46] found larger measurement errors in men as compared with women. A recent article clearly indicated that factors such as pulse arrival time, pulse transit time, systolic pulse transit time and the ratio of areas under the PPG waveform are affected by sex, with men showing a larger effect on the PPG signal.[48] Considering these findings, it appears that sex likely affects

wearable device validity and needs to be accounted for when evaluating PPG-based devices.

Body height

Body height was previously identified as a contributor to larger pulse transit time due to longer distances between the heart and the periphery (ie, the measurement site). Indeed, the latency time between the QRS complex and the PPG peak was found to be larger in taller subjects and was significantly associated with changes in pulse transit time both on the fingers and toes but not on the ears.[44] However, in this study, no associations were found between body height and HR and these findings correspond well with a separate regression analysis,[46] showing body height not to interact with PPG-based HR validity (range: men 159.1–190.0 cm, women 154.4–184.2 cm). In fact, the latter study was the only group that accounted for body height when analysing their findings. However, whether possible delays in pulse transit time were accounted for in the algorithm of the devices tested in the other studies identified by our systematic search remains unknown due to restricted access to algorithms. Thus, we recommend that body height should be considered as a possible factor that may affect HR assessed by PPG. This might be accounted for by including heterogeneous samples of participants (eg, children, adolescents and adults).

Body mass index

In two studies identified by our systematic review, possible associations between BMI and measurement error were assessed but it was concluded that BMI does not affect PPG-based HR accuracy.[50] However, both studies used a rather homogeneous sample (ie, BMI of 20–27 kg/m^2 and 19–33 kg/m^2, respectively). In contrast, in two studies using larger ranges for BMI (ie, 17.2–39.3 kg/m^2 and 17.1–45.0 kg/m^2, respectively) a higher BMI was statistically associated with larger error rates across multiple devices.[40 46] Moreover, a study also found that BMI was correlated with wrist circumference, which may in turn affect the PPG signal.[36] Thus, previous findings provide potential indications for BMI to affect HR measurements assessed by PPG.

Skin tone

The interaction of light with biological tissue can be quite complex and may involve scattering, absorption and/or reflection.[51] For example, it was previously shown that darker skin pigmentation may attenuate the permeability of light wavelengths shorter than 650 nm.[37] The importance of skin tone is underlined by our systematic literature review, showing that skin tone or at least ethnicity was considered as a confounder in 20 studies.[33 35–37 39–41 46 50 52–62] Among these studies, a higher device error was found in participants across several devices or types of activities with darker skin tones assessed by the Fitzpatrick scale,[37 39 46 56 58] while in other studies this was not observed.[33 50 63] Thus, skin tone appears to affect the accuracy of HR readings based on PPG and should also be considered during validity testing.

Criterion measure

The current gold-standard reference method for assessing HR is ECG with 12-lead, which is standard in clinical practice when a full ECG tracing of the cardiac cycle is desired.[64] The largest and most distinct feature in the QRS complex is the R wave, that represents the early ventricular depolarisation and is commonly used for identifying single cardiac cycles. ECG measurements are conducted using dry or wet surface Ag/AgCl electrodes. Wet electrodes include a conductive gel used to decrease the electrode-skin impedance and, thus, increase the signal to noise ratio. However, the conductive gel tends to dry out with time, which potentially affects the data quality.[65] Dry electrodes are an alternative to wet electrodes although not commonly applied with long duration ECG measurements.

Chest strap HR monitors are electronically similar to dry electrodes and some commercially available devices provide beat to beat (RR) intervals for HRV analysis.[66] Chest strap HR monitors are specifically designed to be used with sports participation at various intensities. However, measuring RR intervals during strenuous exercise activities is challenging due to the substantial movement of the torso and occasionally high force impacts which generate motion artefacts in the ECG signal.[67] The validity of estimating RR intervals with commercially available chest strap devices has been investigated in various studies[68–72] and several devices provide RR intervals that demonstrate good to perfect agreement with ECG both in resting conditions and exercise. In online supplemental table 7, we have summarised several validated chest strap devices for measuring RR intervals. The HRV task force guidelines (and the update from 2015) suggest that an independent evaluation of commercially available devices is needed to ensure the validity of the RR interval with HRV analysis.[73 74] However, a required level of agreement for a device to be valid is not specified in these guidelines. Consequently, we suggest that any commercial device (chest strap or ECG measured using either dry or wet electrodes) providing RR intervals, which has been independently validated and demonstrates an excellent agreement with respect to bpm (ie, >95%, see online supplemental table 7), can be used as an appropriate criterion measure for evaluating wearable technologies providing HR to the end-users.

Index measure

In addition to the target population, potential sources of bias originating from the placement of the index device need to be considered. We recommend wearing the index device according to manufacturer's instructions, which should result in standardisation. Nevertheless, the following sections provide a short overview on the most robust factors related to the device placement that may affect the validity of PPG-based HR readings.

Motion artefacts

Previous studies have indicated that consumer wearables are reasonably accurate at resting and moderate steady-state intensities, while the accuracy is typically lower in activities inducing fluctuations in HR.[75] For example, the study by Müller *et al* showed relative higher errors in two activity trackers in a free-living condition compared with a laboratory-based cycling protocol.[62] It is likely that these differences in accuracy are attributed to motion artefacts, which are typically caused by displacement of the PPG sensor over the skin, changes in skin deformation, blood flow dynamics and/or ambient temperature.[76 77] This, in turn, may well manifest as missing or false beats, resulting in invalid HR calculations.[78–80] Even though it is likely that motion artefacts are apparent in every dynamic protocol, only five studies identified by our systematic search specifically reported signal noise originating from movement[37 41 58 81 82] (online supplemental table 6). Thus, protocols used for validity testing of wearables are recommended to include heterogeneous activities or in case the device is intended to be used in a sport or activity-specific setting, conditions similar to the intended setting (ie, providing a common level of movement) should be tested.

Contact pressure

It was previously shown that the waveform of the PPG signal may be affected by the contact force between the sensor and the measurement site and that the waveform of the obtained PPG signal differs depending on the PPG probe contact.[51] The authors further stated that the most accurate PPG signal may be obtained under conditions of transmural pressure, defined as the pressure difference between the inside and outside of blood vessels (ie, the pressure across the wall of the blood vessel). Interestingly, none of the studies identified by our systematic search reported that contact pressure was measured. Future studies should assess whether the validity of HR readings indeed differs between different contact pressures and whether this is related to wearing comfort (ie, ecological validity). Thus, as for now it is recommended to wear the device according to manufacturer instructions during validity testing and to ensure a constant contact pressure, especially in the context of long-term HR monitoring (ie, by repositioning the device periodically).

Ambient light

Light sensitive diodes may also be affected by ambient light. While this has been discussed in few studies identified by our systematic review,[33 37 41 81 83] the magnitude of this effect remains unknown at this stage. In this context, light interferences may be reduced by shading of the interface area site and by electronic filtering (eg, light modulation filtering).[84] Consequently, future studies should address ambient light as a potential source of bias in PPG measurements. Irrespective of this, we believe that potential irritations caused by the ambient conditions may be minimised by correct positioning of the device, as was previously also stated in a topical review.[84]

Ambient temperature

PPG signal quality may also be influenced by the temperature originating both from the environment and changes in skin temperature. While eight studies included in our systematic review reported a controlled laboratory temperature,[34 37 50 53 62 82 85 86] in one study the underestimation of the index device was partially explained by low ambient temperatures of the laboratory assessment (18°C–20°C) compared with that of free-living conditions (30°C–32°C).[62] In addition, it was shown that the error in HR readings obtained from infrared but not green light appeared to be higher in cold (10°C) compared with hot (45°C) conditions.[87] Thus, it seems plausible that ambient temperature may affect the PPG signal quality and should be standardised during laboratory validations and considered as a potential source of bias in free-living conditions.

Testing conditions: laboratory, semifree-living and free-living

Factors that affect the choice of protocol for examining the validity of wearable HR devices, and the validity of the device itself, include types of activities, the intensity of these activities and for how long and frequent these activities are performed. In general, agreement of a device compared with a criterion method during a specific type of activity with a specific intensity is only valid for these conditions. Thus, validity testing programmes of wearable devices may vary in length and complexity and should reflect the intended use of the device. Laboratory based protocols including steady-state activities of varying intensities may be appropriate when examining the basic validity of a device against a criterion, whereas free-living protocols are required when the device is intended to be used in everyday life including sleep.

Furthermore, we recommend to standardise the pretest preparation with a standardised meal replacement to avoid gastric complications during high exercise intensities. Moreover, caffeine should be avoided 12 hours and intense physical activity 48 hours prior to testing. In addition, we recommend a medical screening and to exclude participants using regular medication that affects cardiovascular function (eg, beta-blockers).

Types of activities

Lab based protocols in the studies identified by our systematic search usually included treadmill locomotion[31 32 38–40 53 57 82 88 89] or a combination of treadmill locomotion and ergometer cycling.[33 41 46 52 90–92] In addition, few studies have included activities of daily living (eg, folding laundry and sweeping)[12 45 54 93 94] or resistance exercise[12 50 58 83 94 95] (see online supplemental table 4 for additional information).

HR data measured by consumer wearables were most accurate when assessing locomotor activities that are characterised by repetitive movements (eg, cycling, walking or running) in laboratory settings,[32 40 41 46 52 58] and were less accurate where the movements were inherently more complex, such as resistance exercise and activities of daily living.[50 58] For example, the accuracy of the wearable device was substantially higher during aerobic exercise (92%) as compared with resistance exercises (35%).[50] Similarly, HR measured by non-wrist worn devices (ie, worn in the ear) were relatively accurate during upper and lower body resistance exercises, whereas wrist-worn devices showed poor accuracy. Activities implying upper body movements induced a higher rate of errors and HR drop-outs than endurance exercises (ie, running, walking or cycling) in free-living conditions.[56] As this was not further assessed in the study, it was assumed that this imprecision was due to motion artefacts from the arm and chest movements, as reported during laboratory or semifree-living protocols.[33 41 50 58 95]

Upper body movements cause greater variability in error[41 50 58 91] for wrist-worn devices, probably caused by motion artefacts and variable contact between skin and device, due to variable pressure/contact induced by muscle contractions and changes in blood flow. During upper body work and work involving the arms, muscle and ligament tension in the wrist may interfere with HR detection from capillary blood flow.[50] Thus, devices that rely on HR detection through the skin may be inaccurate if speed or intensity is increased and during activities where skin contact is lost or if an isometric contraction is necessary to perform the activity. Therefore, the specific activities being examined must be clearly considered so the validity of the measurement device in question is aligned with the appropriate/actual use of the device. This is likely an inherited and significant limitation of PPG in general.

Duration and repetitions

The duration of the laboratory protocols performed in the studies identified by our systematic review varied substantially from three to 80 min, with the longest duration observed in semifree-living protocols comprising multiple activities (online supplemental table 4). The length of free-living protocols varied between 2 and 24 hours of continuous monitoring (online supplemental table 4).

Assessing accuracy of HR measurements during steady-state exercise is relevant as consumers often use these devices to monitor HR during continuous exercise sessions or to monitor exercise load and energy expenditure. Steady-state is reached when the HR plateaus during continuous exercise at a submaximal

intensity level, and reflects the balance obtained when the cardiac output is sufficient to transport the oxygen needed to meet the energy cost of the work performed.[96] This usually occurs within the first 2 min of exercise, depending on the change in intensity and fitness level of the participant.[96] However, since HR tends to exhibit a lag compared with the external work performed or the corresponding oxygen cost, protocols should strive for a combination of steady-state activities and those with shorter duration and varying intensities. Indeed, some previous studies have reported lower accuracy when the activity is intermittent with swift changes in exercise intensity (ie, changes in speed of running) or changes in activity that cause changes in wrist movements for PPG wrist-worn devices.[50 58] Since PPG sensors estimate HR by measuring changes in blood flow, the limited blood flow to the wrist at the initiation of exercise might lower the confidence of the predictive algorithms to accurately measure HR (ie, measured heart beats are discarded until the algorithm is confident that it is recording a physiologically plausible value).[97] This must be considered during measurements of HR during activity with rapid changes in intensity and non-steady-state conditions (less than 3 min in duration).

Exercise intensities

Accurate HR readings throughout a wide range of intensities from rest to near maximal is a prerequisite for any consumer device. The laboratory studies reviewed predominantly included multiple intensities ranging from light to very vigorous in their protocols, whereas the measure of intensity varied (eg, speed, watts, metabolic equivalents, % of maximal aerobic capacity) (online supplemental table 4). Semifree-living protocols also included various intensities (online supplemental table 4), whereas the intensity and variability in intensity during free-living is population specific (eg, athletes vs elderly) and cannot be controlled. The accuracy obtained during a free-living protocol also depends on the duration of the measurement and the variability in activities performed (see above). Relatively high measurement errors (10.1%) were observed in a study evaluating the accuracy of a wrist-worn device during a sedentary and light intensity semifree-living protocol,[93] which may be attributable to the freely selection of activities during the testing period. However, this format theoretically mimics everyday activity more effectively than traditional structured activities. It may, therefore, reflect more realistic estimates of validity than a laboratory protocol and may also provide new evidence of how effective the PPG technology is when used in consumer devices.

The studies identified by our systematic review clearly indicated that the accuracy of PPG devices is intensity dependent,[58 83 91 94] with increased accuracy during lower intensity exercise and at rest as compared with vigorous intensity exercise, such as running.[52 57 58 62 83 89 95 98] Conversely, the opposite has also been reported.[32 91] For example, Stahl *et al* observed the highest accuracy (3.06%) during the highest speed tested (9.6 km/hour on the treadmill). One possible explanation is that with increased intensity perfusion is improved, which could decrease the error rate. Consequently, exercise intensities clearly have a profound effect on accuracy of HR readings and should be considered when designing validity protocols in laboratory, semifree-living and free-living conditions.

Processing of index and criterion data

Data processing and reporting is an integral part of validity testing and should be performed with caution. The following items provide recommendations that should be considered in terms of a best practice in the validation process.

Index and criterion synchronisation

Evaluating the validity of consumer wearables measuring HR requires the comparison of two or more time series, which require an equal sampling interval and accurate temporal alignment. The sampling interval of the criterion measure and the wearable devices is most likely not the same and this can be addressed by either interpolation or simple resampling (averaging) of one of the time series. All studies included in the systematic review used resampling to ensure the equal sampling interval. However, out of the 44 included studies only 14 studies[12 34 36 37 50 56 58 62 81 83 90 91 99 100] described the synchronisation process and in three studies[12 56 91] an automated method was performed, whereas in the remaining 11 studies[34 36 37 50 58 59 62 81 83 90 99 100] a manual timestamp correction or visual method was used (online supplemental table 6). The study by Sartor *et al*[12] was the only one included both an automated and visual inspection. Manual correction using time stamps or visual inspection is an option, but the process is time-consuming and potentially error prone. Several methods are available for the automated synchronisation of two independent time series[101 102] and we recommend this approach. The performance of different methods currently available has not been investigated and new methods are continuously being developed. This makes it difficult to propose one single method for the optimal temporal alignment. We recommend that studies use a method that is publicly documented and has been benchmarked with reference to a data set that has been manually edited or generated synthetically.

Data smoothing

Different sources of error may affect the criterion assessment of the RR interval from recordings during both sedentary activities and strenuous exercise. Some errors are related to placement or handling of the device, which can be minimised by the correct application as proposed by the manufacturer. However, ectopic beats (ie, the heart either skips or adds an extra beat) and motion artefacts are errors that are inherent with both chest strap and electrode ECG devices and must be addressed to provide an accurate RR interval with the criterion measure. Only ten studies identified by our systematic literature review described a method to reduce spurious and incorrect HR data[12 34 36 37 56 59 81 82 90 91] (online supplemental table 6). Seven of these studies used an automated method (software), and three studies used a manual method but did not describe this in detail. In the HRV Task Force guideline, it is suggested that manual editing of the RR interval is required for the optimal identification and handling of ectopic beats and motion artefacts.[73 74] Manual editing of long duration recordings is time consuming and requires expert training. However, since the proposal of the HRV guidelines (and the update in 2015) several new methods have been evaluated for the automatic identification and handling of ectopic beats and motion artefacts.[74 103 104] Some of these new methods demonstrate good validity with the assessment of instantaneous HR from RR intervals and should be considered for the validity testing. As with the temporal alignment, currently there is no study available that compares the performance of all the different methods and, therefore, no single method is suggested for the optimal handling of ectopic beats and motion artefacts. We recommend that studies use a method that is publicly documented and has been benchmarked with reference to a data set that has been manually edited or generated synthetically.

Statistical analysis

Since HR is a continuously scaled parameter, the analysis of accuracy should primarily be based on estimation of mean difference or mean relative difference and Bland-Altman limits of agreement (LoA) analysis.[105] The calculated estimates of mean difference and the LoA for the mean difference should always be accompanied with 95% confidence intervals (CIs). The acceptable accuracy expressed as mean difference (bpm) or percentage difference between the criterion measure and the index device may vary and needs to be evaluated individually considering the factors described above. The LoA for the absolute or relative difference are expected to contain 95% of paired differences for each measurement point by the two methods. However, the estimated LoA only apply to the specific study sample and because of sampling error, a new study sample from the same target population might provide different limits. Thus, if accuracy should be compared between different devices (ie, different models and/or manufacturers), it is important to provide the CIs of the LoA and the mean differences. Furthermore, for steady-state activities (in lab and semifree-living conditions) we also recommend that the LoA analysis should be based on both individually averaged mean differences of pairs of HR epochs across the activity duration, and in a repeated measure LoA analysis (multiple paired observations of HR epochs per individual). We acknowledge that validity testing may be performed in order to provide accuracy levels to consumers but also in order to further improve readings of a given device. While not necessarily informative for consumers, in research-related validity testing also proportional or fixed bias may need to be considered. If there is evidence of proportional bias, this suggests that the index device does not agree equally with the criterion through the range of measurements. In this situation, researchers could also use least-products regression as part of the Bland-Altman analysis, as reported by Ludbrook.[106] In case of violations of these assumptions, evaluators could attempt to log-transform index and criterion data or use a non-parametric approach, as described by Bland and Altman.[107]

A correlation coefficient could also be estimated (ie, Pearson's r or concordance correlation coefficient) to provide an additional measure of the relationship between the index and criterion measures, however, the limitations of these measures should be acknowledged as described previously[105] and repeated observations per individual should be taken into account if applicable.

Because HR is obtained in a time series in the wearable consumer device and the criterion measure, the mean difference and the LOAs should be estimated while taking into account multiple observations per individual.[107] We recommend that evaluators check and report on the assumptions for estimating mean difference and LoA. Accordingly, the paired differences in HR from the wearable consumer device and the criterion measure should have an approximately normal distribution, constant variance of the differences between the two, and no proportional bias.[107]

As an additional secondary measure of accuracy, we also recommend reporting the proportion of the evaluated epochs (eg, the exact RR time interval or the averaged HR over a segment of time) of the wearable consumer device that were within the predefined maximum allowed difference and a range of differences of greater and less than the predefined allowed difference.[108] For example, the number of evaluated epochs within ±20 bpm, ±15 bpm, ±10 bpm, ±5 bpm and ±2 bpm. Finally, because some consumer devices may remove data points, for example due to motion artefacts, we recommend reporting the proportion of such missing epochs (total time duration of recorded but missing epochs) of the total epochs recorded. Descriptive data on the study sample, number of paired observations, mean and SD of the HR obtained from the consumer device and the criterion, the mean differences (with SD and standard error), and the mean absolute error and mean absolute percentage error should also be reported.

The within-device precision (ie, reliability) should also be reported based on the data obtained for steady-state activities with a minimum duration of 2 min. To limit the possibility of true biological variation in HR within participants, the within-device precision should be evaluated using the average HR over five seconds separately in each steady-state activity (during rest and exercise) of at least 2 min duration. Furthermore, we suggest that 95% prediction intervals and intraclass correlations with 95% CIs should be calculated to estimate within-device precision according to recommendations.[108]

For a detailed comparison of the statistical analysis in criterion and index devices used in the studies reviewed, please refer to online supplemental table 5).

RECOMMENDED VALIDATION PROTOCOL

Studies aiming to determine the validity of a consumer wearable should be designed to evaluate the device against an accurate criterion measure in a relevant study sample and in conditions that reflect the expected use of the device. Furthermore, the evaluation should be sufficiently described, and the data should be appropriately processed, analysed and reported. Considering the domains presented above, it appears that validation protocols should be carefully designed in order to account at least for the most robust sources of bias. Based on the current state of knowledge, figure 1 provides a graphical matrix of factors that need to be considered when validating PPG-based devices against a gold standard criterion measure. Detailed recommendations and guidelines are provided in online supplemental table 1. In addition, in table 1, we provide a checklist that is intended for planning of the validity protocols. Furthermore, in table 2, a more comprehensive protocol reporting sheet can be found and is intended to be used by both research institutions and manufacturers in order to facilitate standardised and transparent data sharing

DISCUSSION AND FUTURE DIRECTIONS

This expert statement of the INTERLIVE Network aimed to provide recommendations and guidelines for the validity testing of consumer wearables assessing HR by PPG. In this context, considerations for the test preparation, sampling of participants, testing protocols, and activities as well as data handling, analysis and reporting were critically discussed. Based on a systematic literature review as well as our evidence-informed expert opinion, we have suggested a framework for validity testing of PPG-based devices measuring HR.

A rigorous evaluation of validity should be the mutual interest of manufacturers, shareholders, scientific institutions and consumers in order to judge whether a wearable device for assessment of HR is useful and performs with satisfactory accuracy. At present, the decision on whether a validation of a PPG-based wearable complies with medical certifications lies with the manufacturer, inevitably leading to a large heterogeneity in validation protocols. However, new regulations have been put in force on 25 May 2020, requiring all wearables (including devices assessing HR based on PPG) after a transition period of 3 years to follow regulations for medical devices, such as the US

Food and Drug Administration or the CE Marking in Europe.[109] Importantly these regulations are to be adhered to even if the devices are not intended to be used for medical evaluation, risk stratification or patient treatment. Consequently, the present expert statement should be understood as an attempt to foster standards in validations of PPG-based consumer wearables.

The urgent need for standardised validity testing is underlined by the wealth of different protocols identified by our systematic literature review. Interestingly, even though a variety of potential sources of bias were acknowledged in most of these studies, only few have attempted to account for methodological shortcomings in the data analysis. Based on the existing literature, solid evidence exists for artefacts originating from sex, BMI, body height as well factors related to the placement of the device, such as motion artefacts (originating both from the movement itself but also from possible shifts of the device on the skin) and skin tone. Conversely, currently little is known on the effects of cardiovascular diseases and their medical treatment (eg, beta blockers, ACE inhibitors, calcium channel blockers or diuretics), as well as skin damage (eg, scars and tattoos), ambient light and temperature or contact pressure on the measurement of HR by PPG. It is likely that these factors may have profound effects on the validity testing[51] but accounting for this remains challenging. Future studies should focus on the potential sources of bias that stem from both technological as well as population-based characteristics, in order further refine validation protocols.

When considering the proposed recommendations and guidelines, one has to bear in mind that our approach included an initial systematic literature review in order to assess which protocols have previously been used in the scientific literature for validity assessments of PPG-based HR monitors. Thus, we aimed at summarising all studies that have validated numerous index devices against a gold-standard criterion measure and extracted the specifics of these protocols. Consequently, assessing the study quality by means of a risk of bias assessment was not deemed useful as this analysis could only be used to evaluate the quality in respect to the particular outcome of each study (ie, the level of agreement of a certain device compared with a criterion) but it does not provide information on the quality of the validation protocols. Therefore, the potential sources of bias that were indirectly addressed in these studies were aligned with our evidence-informed expert opinions and provided the base for the presented framework.

It is obvious that a best-practice protocol for standardised validations will need to consider interests of both the scientific community and/or customers as well as that of the manufacturers. In that regard, the required investments that has to be made in hardware and software engineering might not be substantial as this is already in place with most manufacturers, but the validity assessment might require employment of additional educated staff and resources for the actual evaluations. Moreover, the resources required to conduct the validity evaluation seem to increase proportionally with the extent of the optimal evaluation. Considering the number of different devices commonly available with many manufacturers, it clearly suggests that feasibility and simplicity is important for proposing a validity evaluation that will be adopted by manufacturers. Consequently, we believe that the present expert statement provides a reasonable base for validity testing by incorporating high scientific standards.

Considering the wealth of new wearables entering the consumer market without prior proof of validity, providing consumers with wearables demonstrating excellent validity seems to be a great opportunity for new companies to conquer a substantial market share. Furthermore, accurate devices will also increase the usability of PPG for diagnostics and therapeutic monitoring. Considering the worldwide increasing prevalence of cardiovascular diseases,[110–112] accuracy will likely become a key criteria for PPG-based wearables. Indeed, prototypes of wrist-worn devices exist that can sense radial artery pulsation and use the data to estimate central aortic pressure.[113] It is likely that wearable devices will soon be capable of extrapolating blood pressure[114] or even blood glucose concentrations through optical sensors,[115] thus, underlining the importance of rigorousness and transparency in evaluating criterion validity. As such, we hope that the provided recommendations and checklists will be deemed useful by both researchers and manufacturers in order to further foster standardised validity testing.

CONCLUSIONS

This expert statement provides an evidence-informed best-practice protocol for the validation of consumer wearables assessing HR by PPG. Our initial systematic literature review underlined a high degree of heterogeneity between previously published methods, with many studies failing to address key sources of bias. Thus, the INTERLIVE Consortium recommends that the proposed validation protocol could be used when considering the validation of any PPG-based consumer wearable assessing HR, in order to overcome the methodological shortcomings highlighted in this statement. Adherence to this validation standard will help ensure a transparent methodological and reporting consistency and facilitate comparison between consumer devices. This will ensure that manufacturers, consumers, healthcare providers and researchers can use this technology safely and to its full potential.

Author affiliations
[1]Institute of Cardiovascular Research and Sports Medicine, Department of Molecular and Cellular Sports Medicine, German Sport University Cologne, Cologne, Germany
[2]Department of Sports Medicine, Norwegian School of Sport Sciences, Oslo, Norway
[3]Department of Sports Science and Clinical Biomechanics, Research Unit for Exercise Epidemiology, Centre of Research in Childhood Health, University of Southern Denmark, Odense, Denmark
[4]Exercise and Health Laboratory, CIPER, Faculdade de Motricidade Humana, Universidade de Lisboa, Lisboa, Portugal
[5]CIDEFES - Centro de Investigação em Desporto, Educação Física e Exercício e Saúde, Universidade Lusófona, Lisboa, Portugal
[6]PROFITH "PROmoting FITness and Health through physical activity" Research Group, Sport and Health University Research Institute (iMUDS), Department of Physical and Sports Education, Faculty of Sport Sciences, University of Granada, Granada, Spain
[7]SFI Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland
[8]School of Public Health, Physiotherapy and Sports Science, University College Dublin, Dublin, Ireland
[9]Exercise and Health Laboratory, CIPER, Faculdade de Motricidade Humana, Universidade de Lisboa, Lisboa, Cruz-Quebrada Dafundo, Portugal
[10]Department of Biosciences and Nutrition, Karolinska Institute, Stockholm, Sweden
[11]Exercise Translational Medicine Centre, the Key Laboratory of Systems Biomedicine, Ministry of Education, and Exercise, Health and Technology Centre, Department of Physical Education, Shanghai Jiao Tong University, Shanghai, China

Twitter William Johnston @Will_Johns10, Ulf Ekelund @Ulf_Ekelund and Moritz Schumann @moritz_schumann

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

Funding JMM is partly funded by Private Stiftung Ewald Marquardt für Wissenschaft und Technik, Kunst und Kultur. UE and JS are partly funded by the Research Council of Norway (249932/F20). ELS is supported by TrygFonden (grant number 310081). PBJ is supported by the Portuguese Foundation for Science and Technology (SFRH/BPD/115977/2016). PMG and FBO are supported by grants from the MINECO/FEDER (DEP2016-79512-R) and from the University of Granada,

Plan Propio de Investigación 2016, Excellence actions: Units of Excellence; Scientific Excellence Unit on Exercise and Health (UCEES); Junta de Andalucía, Consejería de Conocimiento, Investigación y Universidades and European Regional Development Funds (ref. SOMM17/6107/UGR). WJ is partly funded by Science Foundation Ireland (12/RC/2289_P2). AG is supported a European Research Council Grant (grant number 716657). This research was partly funded by Huawei Technologies, Finland.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Jan M Mühlen http://orcid.org/0000-0001-9983-6403
Pedro B Judice http://orcid.org/0000-0003-2791-258X
Pablo Molina-Garcia http://orcid.org/0000-0001-6888-0997
William Johnston http://orcid.org/0000-0003-0525-6577
Luís B Sardinha http://orcid.org/0000-0002-6230-6027
Francisco B Ortega http://orcid.org/0000-0003-2001-1121
Brian Caulfield http://orcid.org/0000-0003-0290-9587
Ulf Ekelund http://orcid.org/0000-0003-2115-9267
Jan Christian Brønd http://orcid.org/0000-0001-6718-3022
Moritz Schumann http://orcid.org/0000-0001-9605-3489

REFERENCES

1 Keith A, Flack M. The Form and Nature of the Muscular Connections between the Primary Divisions of the Vertebrate Heart. *J Anat Physiol* 1907;41:172–89.
2 Fox K, Borer JS, Camm AJ, *et al*. Resting heart rate in cardiovascular disease. *J Am Coll Cardiol* 2007;50:823–30.
3 Åstrand P-O, Ryhming I. A nomogram for calculation of aerobic capacity (physical fitness) from pulse rate during sub-maximal work. *J Appl Physiol* 1954;7:218–21.
4 Kim H-G, Cheon E-J, Bai D-S, *et al*. Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature. *Psychiatry Investig* 2018;15:235–45.
5 Ross R, Blair SN, Arena R, *et al*. Importance of Assessing Cardiorespiratory Fitness in Clinical Practice: A Case for Fitness as a Clinical Vital Sign: A Scientific Statement From the American Heart Association. *Circulation* 2016;134:e653–99.
6 Foster C, Rodriguez-Marroyo JA, de Koning JJ. Monitoring Training Loads: The Past, the Present, and the Future. *Int J Sports Physiol Perform* 2017;12:S22–8.
7 Buchheit M. Monitoring training status with HR measures: do all roads lead to Rome? *Front Physiol* 2014;5:73.
8 Cheung CC, Krahn AD, Andrade JG. The Emerging Role of Wearable Technologies in Detection of Arrhythmia. *Can J Cardiol* 2018;34:1083–7.
9 Pereira T, Tran N, Gadhoumi K, *et al*. Photoplethysmography based atrial fibrillation detection: a review. *NPJ Digit Med* 2020;3:3. doi:10.1038/s41746-019-0207-9
10 Piwek L, Ellis DA, Andrews S, *et al*. The Rise of Consumer Health Wearables: Promises and Barriers. *PLoS Med* 2016;13:e1001953.
11 Raja JM, Elsakr C, Roman S, *et al*. Apple Watch, Wearables, and heart rhythm: where do we stand? *Ann Transl Med* 2019;7:417.
12 Sartor F, Gelissen J, van Dinther R, *et al*. Wrist-worn optical and chest strap heart rate comparison in a heterogeneous sample of healthy individuals and in coronary artery disease patients. *BMC Sports Sci Med Rehabil* 2018;10:10.
13 Turakhia MP, Desai M, Hedlin H, *et al*. Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study. *Am Heart J* 2019;207:66–75.
14 Fortino GG, R.; Galzarano S. *Wearable Computing: From Modeling to Implementation of Wearable Systems Based on Body Sensor Networks*. Wiley, 2018.
15 Bunn JA, Navalta JW, Fountaine CJ, *et al*. Current State of Commercial Wearable Technology in Physical Activity Monitoring 2015-2017. *Int J Exerc Sci* 2018;11:503–15.
16 Sartor F, Papini G, Cox LGE, *et al*. Methodological Shortcomings of Wrist-Worn Heart Rate Monitors Validations. *J Med Internet Res* 2018;20:e10108.
17 Hardey MM. On the body of the consumer: performance-seeking with wearables and health and fitness apps. *Sociol Health Illn* 2019;41:991–1004.
18 Nauman J, Janszky I, Vatten LJ, *et al*. Temporal changes in resting heart rate and deaths from ischemic heart disease. *JAMA* 2011;306:2579–87.
19 John D, Sasaki J, Hickey A, *et al*. ActiGraph™ activity monitors: "the firmware effect". *Med Sci Sports Exerc* 2014;46:834–9.
20 Consumer Technology Association. *ANSI/CTA Standard. Physical Activity Monitoring for Heart Rate. ANSI/CTA-2065*, 2018.
21 Nelson BW, Low CA, Jacobson N, *et al*. Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. *NPJ Digit Med* 2020;3:90.
22 Bassett DR, Rowlands A, Trost SG. Calibration and validation of wearable monitors. *Med Sci Sports Exerc* 2012;44:S32–8.
23 Keadle SK, Lyden KA, Strath SJ, *et al*. A Framework to Evaluate Devices That Assess Physical Behavior. *Exerc Sport Sci Rev* 2019;47:206–14.
24 Freedson P, Bowles HR, Troiano R, *et al*. Assessment of physical activity using wearable monitors: recommendations for monitor calibration and use in the field. *Med Sci Sports Exerc* 2012;44:S1–4.
25 Welk GJ, McClain J, Ainsworth BE. Protocols for evaluating equivalency of accelerometry-based activity monitors. *Med Sci Sports Exerc* 2012;44:S39–49.
26 Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. *AMIA Annu Symp Proc* 2006;2006:359–63.
27 Chatterjee S, Changawala N. Fragmented QRS complex: a novel marker of cardiovascular disease. *Clin Cardiol* 2010;33:68–71.
28 Inui T, Kohno H, Kawasaki Y, *et al*. Use of a Smart Watch for Early Detection of Paroxysmal Atrial Fibrillation: Validation Study. *JMIR Cardio* 2020;4:e14857.
29 Drexler M, Elsner C, Gabelmann V, *et al*. Apple Watch detecting coronary ischaemia during chest pain episodes or an apple a day may keep myocardial infarction away. *Eur Heart J* 2020;41:2224.
30 Hsiu H, Hsu CL, Wu TL. Effects of different contacting pressure on the transfer function between finger photoplethysmographic and radial blood pressure waveforms. *Proc Inst Mech Eng H* 2011;225:575–83.
31 Thomson EA, Nuss K, Comstock A, *et al*. Heart rate measures from the Apple Watch, Fitbit Charge HR 2, and electrocardiogram across different exercise intensities. *J Sports Sci* 2019;37:1411–9.
32 Stahl SE, An H-S, Dinkel DM, *et al*. How accurate are the wrist-based heart rate monitors during walking and running activities? Are they accurate enough? *BMJ Open Sport Exerc Med* 2016;2:e000106.
33 Etiwy M, Akhrass Z, Gillinov L, *et al*. Accuracy of wearable heart rate monitors in cardiac rehabilitation. *Cardiovasc Diagn Ther* 2019;9:262–71.
34 Georgiou KE, Dimov RK, Boyanov NB, *et al*. Feasibility of a New Wearable Device to Estimate Acute Stress in Novices During High-fidelity Surgical Simulation. *Folia Med (Plovdiv)* 2019;61:49–60.
35 Hahnen C, Freeman CG, Haldar N, *et al*. Accuracy of Vital Signs Measurements by a Smartwatch and a Portable Health Device: Validation Study. *JMIR Mhealth Uhealth* 2020;8:e16811.
36 Menghini L, Gianfranchi E, Cellini N, *et al*. Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions. *Psychophysiology* 2019;56:e13441.
37 Dur O, Rhoades C, Ng MS, *et al*. Design Rationale and Performance Evaluation of the Wavelet Health Wristband: Benchtop Validation of a Wrist-Worn Physiological Signal Recorder. *JMIR Mhealth Uhealth* 2018;6:e11040.
38 Coca A, Roberge RJ, Williams WJ, *et al*. Physiological monitoring in firefighter ensembles: wearable plethysmographic sensor vest versus standard equipment. *J Occup Environ Hyg* 2009;7:109–14.
39 Pasadyn SR, Soudan M, Gillinov M, *et al*. Accuracy of commercially available heart rate monitors in athletes: a prospective study. *Cardiovasc Diagn Ther* 2019;9:379–85.
40 Dooley EE, Golaszewski NM, Bartholomew JB. Estimating Accuracy at Exercise Intensities: A Comparative Study of Self-Monitoring Heart Rate and Physical Activity Wearable Devices. *JMIR Mhealth Uhealth* 2017;5:e34.
41 Gillinov S, Etiwy M, Wang R, *et al*. Variable Accuracy of Wearable Heart Rate Monitors during Aerobic Exercise. *Med Sci Sports Exerc* 2017;49:1697–703.
42 MJ L, Zhong WH, Liu YX, *et al*. Sample Size for Assessing Agreement between Two Methods of Measurement by Bland-Altman Method. *Int J Biostat* 2016;12.
43 Carstensen B. *Comparing Clinical Measurement Methods: A Practical Guide*. Wiley, 2010.
44 Allen J, Murray A. Age-related changes in peripheral pulse timing characteristics at the ears, fingers and toes. *J Hum Hypertens* 2002;16:711–7.
45 O'Driscoll R, Turicchi J, Hopkins M, *et al*. The validity of two widely used commercial and research-grade activity monitors, during resting, household and activity behaviours. *Health Technol* 2020;10:637–48.
46 Shcherbina A, Mattsson C, Waggott D, *et al*. Accuracy in Wrist-Worn, Sensor-Based measurements of heart rate and energy expenditure in a diverse cohort. *J Pers Med* 2017;7:3.
47 Chow H-W, Yang C-C. Accuracy of Optical Heart Rate Sensing Technology in Wearable Fitness Trackers for Young and Older Adults: Validation and Comparison Study. *JMIR Mhealth Uhealth* 2020;8:e14707.

48 Dehghanojamahalleh S, Kaya M. Sex-Related Differences in Photoplethysmography Signals Measured From Finger and Toe. *IEEE J Transl Eng Health Med* 2019;7:1–7.
49 Firooz A, Rajabi-Estarabadi A, Zartab H, *et al*. The influence of gender and age on the thickness and echo-density of skin. *Skin Res Technol* 2017;23:13–20.
50 Horton JF, Stergiou P, Fung TS, *et al*. Comparison of Polar M600 Optical Heart Rate and ECG Heart Rate during Exercise. *Med Sci Sports Exerc* 2017;49:2600–7.
51 Tamura T, Chen W. *Seamless Healthcare Monitoring: Advancements in Wearable, Attachable and Invisible Devices*. Springer International Publishing, 2018.
52 Wallen MP, Gomersall SR, Keating SE, *et al*. Accuracy of Heart Rate Watches: Implications for Weight Management. *PLoS One* 2016;11:e0154420.
53 Claes J, Buys R, Avila A, *et al*. Validity of heart rate measurements by the Garmin Forerunner 225 at different walking intensities. *J Med Eng Technol* 2017;41:480–5.
54 Hendrikx J, Ruijs LS, Cox LG, *et al*. Clinical Evaluation of the Measurement Performance of the Philips Health Watch: A Within-Person Comparative Study. *JMIR Mhealth Uhealth* 2017;5:e10.
55 Nelson BW, Allen NB. Accuracy of Consumer Wearable Heart Rate Measurement During an Ecologically Valid 24-Hour Period: Intraindividual Validation Study. *JMIR Mhealth Uhealth* 2019;7:e10828.
56 Hermand E, Cassirame J, Ennequin G, *et al*. Validation of a Photoplethysmographic Heart Rate Monitor: Polar OH1. *Int J Sports Med* 2019;40:462–7.
57 Khushhal A, Nichols S, Evans W, *et al*. Validity and Reliability of the Apple Watch for Measuring Heart Rate During Exercise. *Sports Med Int Open* 2017;1:E206–11.
58 Spierer DK, Rosen Z, Litman LL, *et al*. Validation of photoplethysmography as a method to detect heart rate during rest and exercise. *J Med Eng Technol* 2015;39:264–71.
59 Konstantinou P, Trigeorgi A, Georgiou C, *et al*. Comparing apples and oranges or different types of citrus fruits? Using wearable versus stationary devices to analyze psychophysiological data. *Psychophysiology* 2020;57:e13551.
60 Brazendale K, Decker L, Hunt ET, *et al*. Validity and Wearability of Consumer-based fitness Trackers in free-living children. *Int J Exerc Sci* 2019;12:471–82.
61 Pope ZC, Lee JE, Zeng N, *et al*. Validation of Four Smartwatches in Energy Expenditure and Heart Rate Assessment During Exergaming. *Games Health J* 2019;8:205–12.
62 Müller AM, Wang NX, Yao J, *et al*. Heart Rate Measures From Wrist-Worn Activity Trackers in a Laboratory and Free-Living Setting: Validation Study. *JMIR Mhealth Uhealth* 2019;7:e14120.
63 Sañudo B, De Hoyo M, Muñoz-López A, *et al*. Pilot Study Assessing the Influence of Skin Type on the Heart Rate Measurements Obtained by Photoplethysmography with the Apple Watch. *J Med Syst* 2019;43:195.
64 Bayès de Luna A. *Clinical Electrocardiography: A Textbook*. 4th edn. Wiley-Blackwell, 2012.
65 Satija U, Ramkumar B, Manikandan MS. A Review of Signal Processing Techniques for Electrocardiogram Signal Quality Assessment. *IEEE Rev Biomed Eng* 2018;11:36–52.
66 Giles D, Draper N, Neil W. Validity of the Polar V800 heart rate monitor to measure RR intervals at rest. *Eur J Appl Physiol* 2016;116:563–71.
67 Tobon DP, Jayaraman S, Falk TH. Spectro-Temporal Electrocardiogram Analysis for Noise-Robust Heart Rate and Heart Rate Variability Measurement. *IEEE J Transl Eng Health Med* 2017;5:1–11.
68 Kingsley M, Lewis MJ, Marson RE. Comparison of Polar 810s and an ambulatory ECG system for RR interval measurement during progressive exercise. *Int J Sports Med* 2005;26:39–44.
69 Nunan D, Donovan G, Jakovljevic DG, *et al*. Validity and reliability of short-term heart-rate variability from the Polar S810. *Med Sci Sports Exerc* 2009;41:243–50.
70 Caminal P, Sola F, Gomis P, *et al*. Validity of the Polar V800 monitor for measuring heart rate variability in mountain running route conditions. *Eur J Appl Physiol* 2018;118:669–77.
71 Weippert M, Kumar M, Kreuzfeld S, *et al*. Comparison of three mobile devices for measuring R-R intervals and heart rate variability: Polar S810i, Suunto t6 and an ambulatory ECG system. *Eur J Appl Physiol* 2010;109:779–86.
72 Wallén MB, Hasson D, Theorell T, *et al*. Possibilities and limitations of the Polar RS800 in measuring heart rate variability at rest. *Eur J Appl Physiol* 2012;112:1153–65.
73 force T. Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Task force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. *Eur Heart J* 1996;17:354–81.
74 Sassi R, Cerutti S, Lombardi F, *et al*. Advances in heart rate variability signal analysis: joint position statement by the e-Cardiology ESC Working Group and the European Heart Rhythm Association co-endorsed by the Asia Pacific Heart Rhythm Society. *Europace* 2015;17:1341–53.
75 Bent B, Goldstein BA, Kibbe WA, *et al*. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Digit Med* 2020;3:18.
76 Castaneda D, Esparza A, Ghamari M, *et al*. A review on wearable photoplethysmography sensors and their potential future applications in health care. *Int J Biosens Bioelectron* 2018;4:195–202.
77 Lemay M, Bertschi M, Sola J, *et al*. Chapter 2.3 - Application of Optical Heart Rate Monitoring. In: Sazonov E, Neuman MR, eds. *Wearable Sensors*. Oxford: Academic Press, 2014: 105–29.
78 Zhang Y, Song S, Vullings R, *et al*. Motion Artifact Reduction for Wrist-Worn Photoplethysmograph Sensors Based on Different Wavelengths. *Sensors (Basel)* 2019;19:673.
79 Tăuțan A-M, Young A, Wentink E, *et al*. Characterization and reduction of motion artifacts in photoplethysmographic signals from a wrist-worn device. *Annu Int Conf IEEE Eng Med Biol Soc* 2015;2015:6146-9.
80 Tamura T, Maeda Y, Sekine M, *et al*. Wearable Photoplethysmographic Sensors—Past and Present. *Electronics* 2014;3:282–302.
81 Passler S, Müller N, Senner V. In-Ear Pulse Rate Measurement: A Valid Alternative to Heart Rate Derived from Electrocardiography? *Sensors (Basel)* 2019;19:3641.
82 Zheng Y, Leung B, Sy S, *et al*. A clip-free eyeglasses-based wearable monitoring device for measuring photoplethysmograhic signals. *Annu Int Conf IEEE Eng Med Biol Soc* 2012;2012:5022–5.
83 Jo E, Lewis K, Directo D, *et al*. Validation of Biofeedback Wearables for Photoplethysmographic Heart Rate Tracking. *J Sports Sci Med* 2016;15:540–7.
84 Allen J. Photoplethysmography and its application in clinical physiological measurement. *Physiol Meas* 2007;28:R1–39.
85 Falter M, Budts W, Goetschalckx K, *et al*. Accuracy of Apple Watch Measurements for Heart Rate and Energy Expenditure in Patients With Cardiovascular Disease: Cross-Sectional Study. *JMIR Mhealth Uhealth* 2019;7:e11889.
86 Parak J, Uuskoski M, Machek J, *et al*. Estimating Heart Rate, Energy Expenditure, and Physical Performance With a Wrist Photoplethysmographic Device During Running. *JMIR Mhealth Uhealth* 2017;5:e97.
87 Maeda Y, Sekine M, Tamura T. The advantages of wearable green reflected photoplethysmography. *J Med Syst* 2011;35:829–34.
88 Abt G, Bray J, Benson AC. The validity and inter-device variability of the Apple Watch™ for measuring maximal heart rate. *J Sports Sci* 2018;36:1447–52.
89 Cadmus-Bertram L, Gangnon R, Wirkus EJ, *et al*. The Accuracy of Heart Rate Monitoring by Some Wrist-Worn Activity Trackers. *Ann Intern Med* 2017;166:610–2.
90 Hettiarachchi IT, Hanoun S, Nahavandi D, *et al*. Validation of Polar OH1 optical heart rate sensor for moderate and high intensity physical activities. *PLoS One* 2019;14:e0217288.
91 Parak J, Korhonen I. Evaluation of wearable consumer heart rate monitors based on photopletysmography. *Annu Int Conf IEEE Eng Med Biol Soc* 2014;2014:3670–3.
92 Støve MP, Haucke E, Nymann ML, *et al*. Accuracy of the wearable activity tracker Garmin Forerunner 235 for the assessment of heart rate during rest and activity. *J Sports Sci* 2019;37:895–901.
93 Bai Y, Hibbing P, Mantis C, *et al*. Comparative evaluation of heart rate-based monitors: Apple Watch vs Fitbit Charge HR. *J Sports Sci* 2018;36:1734–41.
94 Reddy RK, Pooni R, Zaharieva DP, *et al*. Accuracy of Wrist-Worn Activity Monitors During Common Daily Physical Activities and Types of Structured Exercise: Evaluation Study. *JMIR Mhealth Uhealth* 2018;6:e10338.
95 Boudreaux BD, Hebert EP, Hollander DB, *et al*. Validity of Wearable Activity Monitors during Cycling and Resistance Exercise. *Med Sci Sports Exerc* 2018;50:624–33.
96 McArdle WKF, Katch V. *Exercise Physiology: Energy, Nutrition and Human Performance*. 8 edn. Lippincott Williams & Wilkins, 2014.
97 Johnson JM. Physical training and the control of skin blood flow. *Med Sci Sports Exerc* 1998;30:382–6.
98 Gorny AW, Liew SJ, Tan CS, *et al*. Fitbit charge HR wireless heart rate monitor: validation study conducted under free-living conditions. *JMIR Mhealth Uhealth* 2017;5:e157.
99 Kroll RR, Boyd JG, Maslove DM. Accuracy of a Wrist-Worn Wearable Device for Monitoring Heart Rates in Hospital Inpatients: A Prospective Observational Study. *J Med Internet Res* 2016;18:e253.
100 Pelizzo G, Guddo A, Puglisi A, *et al*. Accuracy of a Wrist-Worn Heart Rate Sensing Device during Elective Pediatric Surgical Procedures. *Children (Basel)* 2018;5:38.
101 Rhudy M. Time Alignment Techniques for Experimental Sensor Data. *IJCSES* 2014;5:1–14.
102 Coakley KJ, Hale P. Alignment of Noisy Signals. *IEEE Trans Instrum Meas* 2001;50:141–9.
103 Ghaleb FA, Kamat MB, Salleh M, *et al*. Two-stage motion artefact reduction algorithm for electrocardiogram using weighted adaptive noise cancelling and recursive Hampel filter. *PLoS One* 2018;13:e0207176.
104 Luo S, Johnston P. A review of electrocardiogram filtering. *J Electrocardiol* 2010;43:486–96.
105 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–10.
106 Ludbrook J. Statistical techniques for comparing measurers and methods of measurement: a critical review. *Clin Exp Pharmacol Physiol* 2002;29:527–36.
107 Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135–60.
108 Kottner J, Audigé L, Brorson S, *et al*. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J Clin Epidemiol* 2011;64:96–106.
109 Ravizza A, De Maria C, Di Pietro L, *et al*. Comprehensive Review on Current and Future Regulatory Requirements on Wearable Sensors in Preclinical and Clinical Testing. *Front Bioeng Biotechnol* 2019;7:313.

110 Shah AD, Langenberg C, Rapsomaniki E, *et al*. Type 2 diabetes and incidence of cardiovascular diseases: a cohort study in 1·9 million people. *Lancet Diabetes Endocrinol* 2015;3:105–13.
111 Andersson C, Vasan RS. Epidemiology of cardiovascular disease in young individuals. *Nat Rev Cardiol* 2018;15:230–40.
112 Bansilal S, Castellano JM, Fuster V. Global burden of CVD: focus on secondary prevention of cardiovascular disease. *Int J Cardiol* 2015;201(Suppl 1):S1–7.
113 Massoomi MR, Handberg EM. Increasing and Evolving Role of Smart Devices in Modern Medicine. *Eur Cardiol* 2019;14:181–6.
114 Elgendi M, Fletcher R, Liang Y, *et al*. The use of photoplethysmography for assessing hypertension. *NPJ Digit Med* 2019;2:60.
115 Rodin D, Kirby M, Sedogin N, *et al*. Comparative accuracy of optical sensor-based wearable system for non-invasive measurement of blood glucose concentration. *Clin Biochem* 2019;65:15–20.
116 Vandenberk T, Stans J, Mortelmans C, *et al*. Clinical validation of heart rate Apps: mixed-methods evaluation study. *JMIR Mhealth Uhealth* 2017;5:e129.
117 Wang Z, Fu S. Evaluation of a strapless heart rate monitor during simulated flight tasks. *J Occup Environ Hyg* 2016;13:185–92.
118 Chellakumar PJ, Brumfield A, Kunderu K, *et al*. Heart rate variability: comparison among devices with different temporal resolutions. *Physiol Meas* 2005;26:979–86.
119 Engström E, Ottosson E, Wohlfart B, *et al*. Comparison of heart rate measured by polar RS400 and ECG, validity and repeatability. *Adv Physiother* 2012;14:115–22.
120 Gilgen-Ammann R, Schweizer T, Wyss T. RR interval signal quality of a heart rate monitor and an ECG Holter at rest and during exercise. *Eur J Appl Physiol* 2019;119:1525–32.
121 Romagnoli M, Alis R, Guillen J, *et al*. A novel device based on smart textile to control heart's activity during exercise. *Australas Phys Eng Sci Med* 2014;37:377–84.

Online Supplementary Data

Table 1. Proposed validation protocol for validity testing of wearable devices assessing heart rate by photoplethysmography.

Methodological Domains	Methodological Variables	Protocol Considerations	Reporting Considerations
1. Target population	**1.1 Demographic and ethnical characteristics**	Previous studies have indicated that body mass index (BMI), body height, skin tone and sex may affect the validity of wearable devices assessing heart rate (HR) by photoplethysmography (PPG). Therefore, the validation of wearables should include the target sample for a given device, including an equal distribution of men and women of different body height (e.g. by including children, adolescents and adults), BMI and skin tone. Alternatively, manufacturers may decide to assess the validity of a given device in a very specific population (i.e. overweight adults). For this, a homogenous sample should be included.	Report sampling method (e.g. random, convenient etc.) distribution of sex and means & ranges for body height, BMI and skin tone (Fitzpatrick scale)
	1.2 Sample size	For homogenous samples, we recommend a minimum of 45 participants as a rule of thumb [43]. Yet, it is advised to conduct a pilot study to obtain the mean and standard deviation of differences between the wearable consumer device and the criterion measure and consider a pre-defined clinical maximum allowed difference to conduct a prior sample size calculation [42].	Explain how the sample size was selected

2. Criterion measure	**2.1 Reference test**	Chest strap or electrocardiography (ECG) using dry or wet electrodes measuring RR intervals are recommended as a criterion measure.	Report the criterion measure used (model and brand). In case of a chest strap, agreement with respect to beats per minute (bpm) should be reported.
	2.2 Placement	The criterion device should be placed according to manufacturer's instructions.	Report placement of the device (manufacturer's instructions and actual placement)
3. Index device	**3.1 Placement**	The index device should be placed according to manufacturer's instructions.	Report placement of the device (manufacturer's instructions and actual placement)
4. Testing conditions	**4.1 Pre-test preparation**	A standardized meal replacement is suggested to avoid gastric complications during high exercise intensities. Caffeine intake should be avoided 12 hours prior to the measurement. In addition, a medical screening is recommended. Participants using regular medication that affects cardiovascular function (e.g. beta blockers) should be excluded. Participants should refrain from intense physical activity 48 hours prior the validation process.	Pre-test standardization should be reported
	4.2 Laboratory assessment protocol	The purpose of the laboratory protocol is to evaluate the intensity specific accuracy of the wearable with resting, walking, running and biking on a treadmill and cycle ergometer. The protocol should include a wide-range of intensity zones and strive for a combination of steady-state activities and those with shorter duration (including rapid changes in intensity). At least three walking intensities and two running intensities should be evaluated. If biking is included at least	Report the type of activity included with exercise intensity described, preferably relative to aerobic capacity (i.e. % of HR_{max} or VO_{2max}) or in absolute values (i.e. speed/incline or W/rpm).

		three intensities should be evaluated. The choice of intensities (or work rates) needs to consider the characteristics of the population being studied and secondly the setting of which the test is performed. The protocols should assess the accuracy at steady-state (work bouts of 2 to 5 minutes) as well as HR kinetics (transitions and recovery). Examples of different protocols in descending order of validity level: 1. Graded ergometer test with a wide range of exercise intensities reported as % of maximal heart rate (HR_{max}) or maximal oxygen uptake (VO_{2max}) including rest and recovery 2. Graded ergometer test with a wide range of exercise intensities reported in absolute values (i.e. speed/incline, watts (W)/repetitions per minute (rpm)) including rest and recovery 3. Graded ergometer test with a moderate range of exercise intensities reported as % of HR_{max} (or VO_{2max}) including rest and recovery 4. Graded ergometer test with a moderate range of exercise intensities reported in absolute values (i.e. speed/incline, W/rpm) including rest and recovery 5. Graded ergometer test with a low range of exercise intensities reported as % of HR_{max} (or VO_{2max}) including rest and recovery 6. Graded ergometer test with a low range of exercise intensities reported in absolute values (i.e. Speed/ incline, W/rpm) including rest and recovery Pre-determination of HR_{max} (or VO_{2max}) that allows for assessing intensities relative to the participant's fitness level are likely to produce more precise intensity estimates but are more time consuming and may be perceived as more	

		invasive for the participant and are therefore not always feasible. In that case, absolute exercise intensities should be used.	
	4.3 Semi-free-living (sport-specific) assessment protocol	The purpose of this evaluation is to assess the accuracy of the index device with different activities that are executed in an environment that is true to the nature of the activity. The duration of the activity should be sufficient to include intensities which commonly describe the inherent nature of the activity (continuous or intermittent). Evaluating accuracy of a devices with an intermittent activity like soccer (sport-specific) require the activity to be performed on a standard playing field (artificial or natural grass) and to include several players that will ensure a sufficient game intensity. If these prerequisites are met, it is sufficient to use a measurement duration of 15-20 minutes. When evaluating the accuracy with more continuous activities (running, walking, biking, swimming), the execution of the activity should include a minimum of three different intensities (approximately 40 %, 60 %, 80 % of HR_{max}) and each of intensity should have a duration of minimum 4 - 5 minutes. Evaluating the activities that are common with domestic behaviour (doing laundry, gardening, home construction, office work) require a duration of at least 15 - 20 minutes.	Description of the activity included and the duration
	4.4 Free-living assessment protocol	The measurement protocol for evaluating the 24 hour HR accuracy is trivial and only requires the included subjects to wear the index and criterion device for a duration of 24 hours during the subject's normal daily living. Subjects not presenting HR data above 40 % of HR_{max} should be excluded from the evaluation. Similarly, recordings	Report the duration of testing.

		missing more than 5 % of the data in either index or criterion should also be excluded.	
5. Processing	**5.1 Criterion measure processing**	An automated method must be applied with the RR intervals to account for motion artefacts and ectopic beats.	The method used for error correction and data smoothing.
	5.2 Index measure processing	No post processing of the end-user HR data is allowed, although the resampling into a window size of 5 seconds is allowed.	
	5.3 Epochs for analysis/window size	The criterion measure must be sampled using the same window size (epoch) as available with the index measure. The window size should be 5 seconds or shorter.	
	5.4 Index and criterion synchronisation	An automated method for synchronizing the criterion and index measure must be used (cross correlation or similar methods).	The method used.

6. Statistical analysis	**6.1 Statistical tests**	Mean difference or mean relative difference and Bland-Altman limits of agreement (LoA) analysis should be performed. To be able to compare evaluations between different devices, we recommend as a minimum that analysis should be based on 5 second windows. A repeated measure LoA analysis (multiple paired observations of HR epochs per individual) should be used in non-steady-state conditions, however, we also recommend, for the steady-state activities (in lab and semi-free-living conditions), that the LoA analysis should be based on both individually averaged mean differences of pairs of HR epochs across the activity duration. The within-device precision should be evaluated by comparing the within-person variability in average HR over 5 second windows, separately for steady-state activities (during rest and exercise) of at least 2 minutes duration conducted in the lab.	Descriptive data on N of paired observations, mean and standard deviation (SD) of the HR obtained from the consumer device and the criterion, the mean differences (with SD and standard error), and LoA with 95 % confidence intervals (CI). Report that the assumptions for LoA analysis has been checked and dealt with appropriately. The mean absolute error and mean absolute percentage error should also be reported for each steady-state intensity. For the 24 hour evaluation, the mean absolute error and LoA must be reported with all data and in the three domains <100bpm, >=100bpm & <140bpm and >=140bpm. We recommend that 95 % prediction intervals and intra class correlation with 95 % CI should be calculated to estimate within-device precision [108].

Table 2. Search terms used in Embase, Web of Science, and PubMed databases.

Embase	Web of Science	PubMed
Index device	**Index device**	**Index device**
('wearable electronic devices'/exp OR wearable electronic devices' OR wearable* OR watch* OR smartwatch* OR (('smart'/exp OR 'smart') AND watch*) OR (('smart'/exp OR smart) AND band*) OR (('smart'/exp OR smart) AND bracelet*)	ALL FIELDS: (wearable* OR smartwatch* OR "smart watch" OR "smart watches" OR watch* OR (smart AND band*) OR (smart AND bracelet*))	("Wearable Electronic Devices"[Mesh] OR wearable* OR smartwatch* OR watch* OR (smart AND watch*) OR (smart AND band*) OR (smart AND bracelet*)
Outcome	**Outcome**	**Outcome**
AND (('heart'/exp OR heart) AND rate OR 'pulse'/exp OR pulse) AND rate	AND ALL FIELDS (heart AND rate*) OR (pulse AND rate*)	AND ((heart AND rate*) OR pulse AND rate*))
Study design	**Study design**	**Study design**
AND ('reproducibility of results'/exp OR 'reproducibility of results' OR 'validity'/exp OR 'validity' OR 'validation'/exp OR 'validation' OR validate OR 'comparison'/exp OR 'comparison' OR 'reliability'/exp OR 'reliability' OR reliable))	AND ALL FIELDS (validity OR validation OR validate OR comparison OR reliability OR reliable)	AND ("Reproducibility of Results"[Mesh] OR validity OR validation OR validate OR comparison OR reliability OR reliable)

Table 3. Summary of populations used in the studies identified by the systematic literature review.

N	Author (year)	Number of participants	Age (mean ± SD and range)	BMI (mean ± SD and range)	Sex distribution	Skin tone assessment	Wrist circumference (mean ± SD or range)	Preparatory actions	Measurement site
1	Abt (2018)[88]	15	32 ± 10	ND	♂8/♀7	ND	ND	ND	Left and right wrist
2	Bai (2018)[93]	41	32 ± 11 (19-60)	24.7 ± 4.0 (18.5-37.6)	♂23/♀18	ND	ND	ND	Left wrist
3	Boudreaux (2018)[95]	50	♂22 ± 3 ♀23 ± 3 (18-35)	♂27.1 ± 3.6 ♀ 25.8 ± 4.8	♂22/♀28	ND	ND	ND	Left and right wrist and ear
4	Brazendale (2019)[60]	Study 1: 19 Study 2: 20	Study 1: 8 ± 2 Study 2: 9 ± 2	ND	Study 1: ♀46% Study 2: ♀50%	Ethnicity	ND	ND	Non-dominant wrist
5	Cadmus-Bertram (2017)[89]	40	49 ± 10 (30-65)	25.1 ± 3.9	♂20/♀20	ND	ND	ND	Left and right wrist
6	Claes (2017)[53]	12	28 ± 5 (20-40)	22.1 ± 3.5	♂6/♀6	Ethnicity	ND	ND	Left forearm
7	Coca (2010)[38]	10	27 ± 7 (21-39)	25.1 ± 5.7	♂8/♀2	ND	ND	Health screen	Rib cage
8	Dooley (2017)[40]	62	23 ± 4 (18-38)	24.6 ± 4.8 (17.1-45.0)	♂26/♀36	Ethnicity	ND	Caffeine and nutrition restriction	Left or right wrist
9	Dur (2018)[37]	35	25 ± 4	ND	♂19/♀16	Fitzpatrick scale	ND	ND	Left and right wrist
10	Etiwy (2019)[33]	80	62 ± 13	29.0 ± 5.5	♂65/♀15	Ethnicity	Right 18 ± 1.6 Left 18 ± 1.6	Medications	Left and right wrist
11	Falter (2019)[85]	40	62 ± 15	27.0 ± 5.0	♂32/♀8	ND	ND	Diagnosis, smoking	Left wrist
12	Georgiou (2019)[34]	21	23-26	18.5-24.9	Only male	ND	ND	ND	Non-dominant hand (wrist)

13	Gillinov (2017)[41]	50	38 ± 12 (21-64)	25.0 ± 3.5 (19-33)	♂23/♀27	Ethnicity	Right 16.0 ± 1.4 Left 16.0 ± 1.4	ND	Forearm and left or right wrist
14	Gorny (2017)[98]	10	25 ± 4 (18-65)	22.9 ± 3.8	♂7/♀3	ND	ND	Nutrition restriction	Non-dominant hand (wrist)
15	Hahnen (2020)[35]	85	53 ± 21	28.0 ± 7.0	♂49/♀36	Ethnicity	ND	Medication	Wrist
16	Hendrikx (2017)[54]	29	41 ± 14	25.1 ± 3.1 (20.4-31.5)	♂14/♀15	Fitzpatrick scale	ND	Restrictions before test (exercise, nutrition)	Wrist
17	Hermand (2019)[56]	70	20 ± 6	ND	♂56/♀14	Fitzpatrick scale	ND	ND	Upper arm
18	Hettiarachchi (2019)[90]	24	28 ± 6 (21-38)	♂24.4 ± 3.26 ♀20.9 ± 4.57 (16.3–33.3)	♂12/♀12	ND	ND	ND	Forearm and upper arm and temple
19	Horton (2017)[50]	36	41 ± 10 (18-55)	♂24.3 ± 2.3 ♀22.3 ± 2.0 (20.0-27.0)	♂18/♀18	Fitzpatrick scale	Right ♂17.0 ± 1.0 ♀15.1 ± 0.8 Left ♂16.9 ± 1.0 ♀15.0 ± 0.8	ND	Left wrist
20	Jo (2016)[83]	24	25 ± 2	ND	♂12/♀12	ND	ND	ND	Left and right wrist
21	Khushhal (2017)[57]	29	31 ± 7	26.1 ± 2.9	Only male	Ethnicity	ND	Caffeine and nutrition restriction	Left and right wrist
22	Konstantinou (2020)[59]	43	21 ± 4 (18-38)	ND	♂6/♀37	Ethnicity	ND	ND	Non-dominant wrist
23	Kroll (2016)[99]	50	64	ND	♂26/♀24	ND	ND	Diagnosis	Wrist
24	Menghini (2019)[36]	40	30 ± 13 (18-60)	23 ± 4	♂21/♀19	Von Luschan's scale	ND	Restrictions before test (exercise, nutrition)	Non-dominant wrist
25	Müller (2019)[62]	57	31 ± 10 (21-50)	65% with 18.5-23.0	♂29/♀26	Ethnicity	ND	Caffeine and nutrition restriction	Left and right wrist

26	Nelson (2019)[55]	1	29	21	Only male	Fitzpatrick scale	Right 7.0 Left 6.5	ND	Left and right wrist
27	O'Driscoll (2019)[45]	59	44 ± 14 (22-73)	ND	♂18/♀41	ND	ND	Caffeine and nutrition restriction	Non-dominant wrist
28	Parak (2014)[91]	21	31 ± 11	ND	♂15/♀6	ND	ND	Non smokers	Forearm, wrist
29	Parak (2017)[86]	24	36 ± 8 (18-55)	22.7 ± 1.9 (18.0-30.0)	♂13/♀11	ND	ND	ND	Wrist
30	Pasadyn (2019)[39]	50	30 ± 9 (18-56)	22.8 ± 2.4 (18.5-28.3)	♂34/♀16	Ethnicity	Right 16.3 ± 1.1 Left 16.2 ± 1.1	ND	Left and right wrist
31	Passler (2019)[81]	20	22 ± 2	69.6 ± 11.0	♂14/♀6	ND	ND	ND	In-ear
32	Pelizzo (2018)[100]	30	8 ± 3	20.5 ± 5.0	♂16/♀14	ND	ND	ND	Arm
33	Pope (2019)[61]	21	25 ± 4 (18-35)	≤18.5	♂7/♀14	Ethnicity	ND	Medication restriction	Left and right wrist
34	Reddy (2018)[94]	20	28 ± 6	22.5 ± 2.3	♂9/♀11	Ethnicity	15.6 ± 2.0	ND	Left and right wrist
35	Sartor (2018)[12]	199	38 ± 7	25.8 ± 3.0	♂84/♀115	Fitzpatrick scale	ND	ND	Wrist
36	Shcherbina (2017)[46]	60	38 ± 11	♂24.9 ± 3.5 ♀22.4 ± 3.3 (17.2-39.3)	♂29/♀31	Von Luschan's chromatic scale + Fitzpatrick scale	♂17.3 ± 1.1 (16-21) ♀15.4 ± 1.3 (13.5-17.5)	ND	Right and left wrist anterior and posterior
37	Spierer (2015)[58]	50	28 ± 10	ND	♂27/♀20	Fitzpatrick scale	ND	ND	Left and right wrist
38	Stahl (2016)[32]	50	♂27 ± 6 ♀24 ± 45 (19-43)	♂25.4 ± 2.6(19.7-30.7) ♀23.4 ± 3.5 (17.7-31.9)	♂32/♀18	ND	ND	ND	Forearm and left/right wrist

39	Støve (2019)[92]	29	29 ± 9 (18-51)	23.7 ± 2.2 (19.9-28.4)	♂17/♀12	ND	ND	Caffeine, smoking, nutrition and medication restriction	Left wrist
40	Thomson (2019)[31]	30	24 ± 3	22.8 ± 2.2	♂15/♀15	ND	ND	ND	Left and right wrist
41	Vandenberk (2017)[116]	225	75 ± 14	ND	♂105/♀120	ND	ND	ND	Left and right index and middle finger
42	Wallen (2016)[52]	22	24 ± 6	ND	♂11/♀11	Fitzpatrick scale	ND	ND	Left and right arm (wrist)
43	Wang (2016)[117]	10	39 ± 8	ND	Only male	Ethnicity	ND	ND	Left wrist
44	Zheng (2012)[82]	10	27 ± 4	ND	ND	ND	ND	ND	Nose bridge, right index finger, right earlobe

Abbreviations. SD: standard deviation; BMI: body mass index; ND: not disclosed.

Table 4. Summary of protocols used in the studies identified by the systematic literature review.

N	Author (year)	Lab, semi-lab or free-living	Types of activities	Duration/repetitions	Intensities
1	Abt (2018)[88]	Lab	Treadmill walking/running	ND	Graded exercise to exhaustion
2	Bai (2018)[93]	Lab and semi-free-living	Sedentary activity, treadmill walking/running and simulated free-living activities (folding laundry, sweeping, moving light boxes, stretching, slow walking)	80 minutes protocol (20 minutes of sedentary activity, 60 minutes PA)	ND (self-selected pace on treadmill)
3	Boudreaux (2018)[95]	Lab and semi-free-living	Cycling, strength training exercises (2 upper body: chest press, latissimus dorsi pulldown, 2 lower body: leg extension, leg curl)	Cycling: 2 minute stages at 50 rpm, beginning at 300 kpm/minute and increasing by 150 kpm/minute until exhaustion. 3 sets of 4 exercises at 10 RM	Until exhaustion
4	Brazendale (2019)[60]	Free-living	A variety of activities that consisted of staff-led structured games (e.g. tag, basketball) and free-play opportunities	2*2 hour daily segments for 14 days	Sedentary to vigorous
5	Cadmus-Bertram (2017)[89]	Lab	Treadmill walking/running	10 minutes	65% of HR_{max}
6	Claes (2017)[53]	Lab	Treadmill walking/running	3*10 minutes	Moderate to high intensity (4, 6 and >7 METs)
7	Coca (2010)[38]	Lab	Treadmill walking/running	20 minutes	50% of VO_{2max}
8	Dooley (2017)[40]	Lab	Treadmill walking/running	4*4 minute stages	Light (2.5 mph), moderate (3.5 mph), and vigorous (5.5 mph)
9	Dur (2018)[37]	Lab	Sitting	ND	Only resting HR
10	Etiwy (2019)[33]	Lab	Treadmill walking/running and cycling	7 minutes	Steady-state exercise at 50-70% of HR reserve
11	Falter (2019)[85]	Lab	Cycling	ND	Light to vigorous (graded exercise to exhaustion)

12	Georgiou (2019)[34]	Lab	Leisurely reading, basic surgical skills module	10 minutes reading, 9 minutes basic skills exercise	ND
13	Gillinov (2017)[41]	Lab	Treadmill walking/running, cycling and elliptical exercising	24 minutes (3*1.5 minute stages per ergometer)	Light, moderate, and vigorous intensity (2-10 METs)
14	Gorny (2017)[98]	Free-living	Participants were encouraged to continue pursuing their usual activities	1 month	ND
15	Hahnen (2020)[35]	Lab	Sitting	ND	Only resting HR
16	Hendrikx (2017)[54]	Lab, semi-free and free-living	Treadmill running/walking, cycling, outdoor walking and cycling, cross-trainer, household activities	Lab/semi-free: 3 minutes for each activity, separated with 3 minutes rest. Free-living: 3 days	Low to moderate: Treadmill (3-4.5 km/h, 0-5%), ergometer bike (60 rpm), cross trainer (60W)
17	Hermand (2019)[56]	Free-living	Running, biking and walking performed on various terrains (flats, hills and downhills). Tennis, CrossFit and soccer were performed on flat ground.	Recordings were started at rest before the start of exercise and terminated after a short recovery time. In all, 390 hours and 38 minutes of recordings were analysed, distributed across 233 sessions.	A wide HR spectrum from low to high
18	Hettiarachchi (2019)[90]	Lab	Treadmill walking/running and cycling	9+9+6 minutes	Light to vigorous
19	Horton (2017)[50]	Lab and semi-free-living	Cycling, circuit weight training (shoulder shrugs, squats, bicep curls, and lunges)	Total 76 minutes. Participants performed 7*3 minute intervals in a pyramid fashion. Each strength exercise was performed for 30 seconds with no rest between exercises.	Walking speed was 4.0 km/h and jogging speed was 8.0 km/h. The running speed was selected by each subject based on recent 5 km race pace.
20	Jo (2016)[83]	Lab and semi-free-living	Cycling, walking/running, strength training: free-weight arm raises and lunges, and isometric plank	Total 77 minutes. Initial rest period (supine) of 15 minutes, 5 minute bouts activity. 12 repetitions of resistance exercises.	Low (60 W) to intense (120 W) cycling. Walking (3.0-3.5 mph speed), jog (4.0-5.0 mph), run (5.5-7.0 mph).

21	Khushhal (2017)[57]	Lab	Treadmill walking/running	3*5 minutes	Light to vigorous (4, 7 and 10 km/h)
22	Konstantinou (2020)[59]	Lab	Cold pressor pain task	ND	ND
23	Kroll (2016)[99]	Free-living/Clinical setting	In-patients monitored bedside (hospital)	24 hours	ND
24	Menghini (2019)[36]	Lab	Seated paced breathing, orthostatic test, walking, keyboard typing, Stroop test, speech test, public speech, speech recovery	30 minutes (3 minutes each exercise)	ND
25	Müller (2019)[62]	Lab and free-living	Cycling	4*5 minute bouts, free-living the next day	45%-75% of HR_{max}
26	Nelson (2019)[55]	Free-living	Walking, treadmill running; activities of daily living (cleaning, brushing teeth, and cooking) and sleeping	24 hours	ND
27	O'Driscoll (2019)[45]	Lab and semi-free-living	Walking, running, cycling, sedentary and household tasks (folding and sweeping tasks)	7*5 minute bouts of sitting, standing, treadmill walking/running. 3 minutes rest. 2*5 minutes cycling, 3 minutes rest, 2*5 minutes household tasks.	Low to moderate/vigorous (walking 4 km/h, 0-5% incline), running (6-8 km/h, 0-5% incline)
28	Parak (2014)[91]	Lab	Treadmill walking/running and cycling	30 minutes exercise, total protocol 47 minutes	Low to high (3-11 km/h, various incline)
29	Parak (2017)[86]	Lab and semi-free-living	Outdoor and treadmill running	Outdoor: self-determined pace for at least 20 minutes. Indoor: 8-10*3 minute stages.	Outdoor: moderate to vigorous subjectively assessed intensity, and to run 5 km. Indoor: High to exhaustion.
30	Pasadyn (2019)[39]	Lab	Treadmill walking/running	6*2 minute stages	Light to vigorous, graded exercise (4-9 mph)

31	Passler (2019)[81]	Lab	Cycling	20 minutes	Light to vigorous (graded exercise to exhaustion)
32	Pelizzo (2018)[100]	Free-living/Clinical setting	Monitored during surgery	ND	ND
33	Pope (2019)[61]	Semi-free	Exergaming (PA videogames)	20 minutes	ND
34	Reddy (2018)[94]	Lab and semi-free-living	Cycling or treadmill running, circuit free-weight training (arm raises, resisted lunges, and isometric plank) and 6 activities of daily living	2 sets of 8 RM. 6*ADLs (3 minutes in duration). 5*2 minute HIIT	Graded exercise to exhaustion. HIIT at a high intensity (60 rpm), at a power output corresponding to approximately 80% of their peak power output.
35	Sartor (2018)[12]	Lab and semi-free-living	Walking, running (indoor and outdoor), cycling (indoor and outdoor), gym (rowing, stepping, group training), household, and sedentary activities	Lab activities lasted 3 minutes. Outdoor and group fitness activities lasted about 1 hour.	Light to vigorous. Treadmill locomotion 3-16 km/h, 0-10% inclination, cycling 50-200 W or self-paced (outdoor and gym activities).
36	Shcherbina (2017)[46]	Lab	Treadmill walking/running and cycling	5 minute bouts. Total approx. 40 minutes	Light to vigorous (Treadmill, 3-9 mph, cycling 50-225 W)
37	Spierer (2015)[58]	Lab and semi-free-living	Treadmill walking/jogging, elliptical exercise, stair climbing, stationary cycling and light weightlifting	7*6 minute exercise bouts, biceps curl with barbell, 1 kg for women and 2 kg for men	Exercise intensity during all activities apart from light weightlifting was self-selected. Each participant was asked to find a pace that allowed them to endure that level of activity for a minimum of 6 minutes.
38	Stahl (2016)[32]	Lab	Treadmill walking/running	5*5 minutes at each speed	Light to vigorous, graded exercise (3.2-9.6 km/h)
39	Støve (2019)[92]	Lab	Treadmill walking/running and cycling	3*3 minutes cycling and 3*3 minutes walking/running	Submaximal to near-maximal exercise (50, 100 and 150 W cycling and 4.8, 8.7 and 12.1 km/h walking/running)
40	Thomson (2019)[31]	Lab	Treadmill walking/running	2-12 minutes (3 minute stages)	Light to vigorous (graded exercise to exhaustion)

41	Vandenberk (2017)[116]	Lab	Cycling	5 minutes	Light to vigorous (graded exercise to HR_{max})
42	Wallen (2016)[52]	Lab	Treadmill walking/running and cycling	58 minutes total, 3*5 minutes cycling, 6*3 minutes stepping	70-80% of HR_{max}
43	Wang (2016)[117]	Semi-free-living	Simulated flight in flight simulator	ND	ND
44	Zheng (2012)[82]	Lab	Treadmill walking	1 minute	Light (slow walking)

Abbreviations. ND: not disclosed; PA: physical activity; rpm: repetitions per minute; kpm: keystrokes per minute; RM: repetition maximum; HR_{max}: maximal heart rate; METs: metabolic equivalents; VO_{2max}: maximal oxygen uptake; mph: miles per hour; HR: heart rate; W: Watt; ADLs: activities of daily living; HIIT: high intensity interval training.

Table 5. Summary of index and criterion measures used in the studies identified by the systematic literature review.

N	Study	Index device	Criterion measure	Statistical comparison
1	Abt (2018)[88]	Apple Watch™ (watchOS 2.0.1) on each wrist (right and left). HR data were recorded every 5 second on each watch using the "Workout" app.	A Polar T31™ chest strapped HR monitor	Pearson correlation, ICC, Cohen's d, standardised mean bias, and standardised typical error of the estimate
2	Bai (2018)[93]	Apple Watch 1 and Fitbit Charge HR, both fitted on left wrist. The applications for the consumer monitors were initialized to incorporate the participant's demographic and anthropometric information.	Polar chest strap placed just below chest muscles and firmly against the skin. The Oxycon Mobile 5.0 incorporates HR telemetry to record the minute by minute Polar belt HR data as part of its output.	B&A, Pearson correlation, mean percent errors, MAPE, RMSE, equivalence testing
3	Boudreaux (2018)[95]	8 wearable devices (6 wrist-worn, randomized placement, 3 devices on each wrist; 1 chest-worn; one ear-worn) simultaneously:- Apple Watch Series 2 (Apple Inc), Fitbit Blaze (Fitbit Inc), Fitbit Charge 2 (Fitbit Inc), Polar H7 chest strap (Polar Electro), Polar A360 (Polar Electro), Garmin Vivosmart HR (Garmin International Inc), TomTom Touch (TomTom), Bose SoundSport Pulse headphones (Bose Corporation).	6-lead ECG (Quinton 4500, Milwaukee, WI)	B&A, ICC, paired t-test, MAPE
4	Brazendale (2019)[60]	Fitbit Charge HR© to wear on their non-dominant wrist, and a Polar H7© watch on their dominant wrist.	Polar H7© (Polar Electro Inc., Lake Success, NY, USA) telemetry chest strap	Pearson correlation, B&A, MAPE
5	Cadmus-Bertram (2017)[89]	Fitbit Charge (Fitbit), Fitbit Surge (Fitbit), Basis Peak (Basis) and Mio Fuse (Mio Global). All wrist-worn.	ECG	B&A, repeated measures mixed model

6	Claes (2017)[53]	Garmin Forerunner 225 (Garmin International, Kansas City, MO), programmed with the participants' sex, age, weight and height and was fitted on the left forearm.	3-lead ECG (Zensor VR, Intelesens Ltd, Belfast, UK). Attached on the chest with the studded attachment electrode placed directly under the left side of the rib cage and the two 2-electrodes placed on both processus coracoideus at the level of the shoulder.	Pearson correlation, RMSE, B&A, paired t-test
7	Coca (2010)[38]	LifeShirt (VivoMetrics, Ventura, Calif.). Central and peripheral physiological sensors included in wearable plethysmograph sensor vest.	3 ECG electrodes placed at the upper left and upper right anterior chest wall and distal left lateral abdominal wall. (VIASYS/SensorMedics, Yorba Linda, Calif).	Bootstrap estimates
8	Dooley (2017)[40]	3 wrist-worn wearables: Apple Watch, Fitbit Charge HR, and Garmin Forerunner 225.	Polar T31 transmitter monitor worn around the chest and transmits real-time HR of the user to a wristwatch ECG.	B&A, MAPE, 2 way repeated measures analysis of variance
9	Dur (2018)[37]	Wavelet wristband. The LEDs fire at a rate configurable between 20 and 95 Hz driven by a sub millisecond resolution low-jitter external clock signal. For this validation study, light sensor data were collected at 86 Hz.	BIOPAC MP36 system (BIOPAC, Goleta, CA, USA). ECG (LEAD110A and ECG100C, BIOPAC, Goleta, CA, USA) was acquired at a rate of 2000 Hz while the subject was at rest in a seated position.	Pearson correlation, B&A
10	Etiwy (2019)[33]	Fitbit Blaze (Fitbit Inc., San Francisco, CA, USA), Apple Watch (Apple Inc., Cupertino, CA, USA), Garmin Forerunner 235 (Garmin Inc., Olathe, KS, USA), TomTom Spark Cardio (TomTom, Inc., Amsterdam, Netherlands). Wrist-worn monitors were affixed securely above the ulnar styloid. Participants were randomly assigned to wear 2 different wrist-worn HR monitors, 1 on each wrist.	3 lead ECG (Mason-Likar electrode placement of torso-mounted limb leads).	B&A, CCC, repeated measures mixed model
11	Falter (2019)[85]	Apple Watch Sport 42 mm (Apple Inc), left wrist	12-lead ECG (Cardiosoft, General Electric Company)	B&A, CCC

12	Georgiou (2019)[34]	Empatica E4 wristband (E4WB) (Empatica S.r.l, Italy) on their non-dominant hand	3-lead ambulatory Holter ECG rhythm monitoring (HM) and electrodes were positioned in predetermined thorax positions (ELA Medical - Syneflash MMC-24-hour Rhythm - Synescope ELA Medica, France).	Pearson correlation, B&A
13	Gillinov (2017)[41]	Forearm monitor (Scosche Rhythm+), and two randomly assigned wrist-worn HR monitors (Apple Watch, Fitbit Blaze, Garmin Forerunner 235, and TomTom Spark Cardio), 1 on each wrist.	Chest strap monitor (Polar H7)	B&A, CCC, absolute percentage differences, Repeated-measures mixed model ANOVA
14	Gorny (2017)[98]	Fitbit Charge HR (Fitbit) tracker to be worn on the non-dominant hand. Fitbit measures were accessed at 1 minute intervals.	Polar H6 HR (Polar Electro Oy, Kempele, Finland) worn across the chest, while Polar readings were available for 10 second intervals. To record the Polar H6 HR monitor (Polar) data, these participants were provided with an Actigraph GT3X+ logger (Actigraph) on Bluetooth receiver mode set to sample measures at 10 second intervals and worn on the same wrist as the Fitbit device.	B&A, ICC
15	Hahnen (2020)[35]	The Everlast TR10 smartwatch. Wrist-worn. To measure HR, the right index finger needs to be placed beneath the cap on top, the right thumb on the electrode on the front, and the right middle finger on the electrode on the back of the device. Measure time approx. 30 seconds. Require the input of sex, date of birth, height and weight.	Cardiocap/5 (Datex-Ohmeda) hospital-grade vital signs monitor (HR can be measured using ECG or can be derived from the SpO2, PPG-driven). Everlast smartwatch and BodiMetrics tricorder were prepared according to their manufacturers' guidelines.	Pearson correlation, B&A, mean absolute difference

16	Hendrikx (2017)[54]	Philips health watch, wrist-worn, 1 Hz sampling rate, displays real-time HR. 1 minute average values for HR over 1 minute, are logged in internal memory and transmitted via Bluetooth to a phone running the companion app for 24/7 monitoring.	The Actiwave Cardio (CamNtech, Cambridge, UK), a single-channel ECG waveform recorder that participants wore (only) during the laboratory protocol and it reported HR at a frequency of 1 Hz.	Equivalence tests of paired means
17	Hermand (2019)[56]	Polar OH1 strapped around the upper arm, firmly enough to remain in place but not enough to obstruct blood flow. Recordings for both were started at rest before the exercise start and terminated after a short recovery time.	Polar H7 chest belt paired with a Polar M400 watch	B&A, CCC
18	Hettiarachchi (2019)[90]	Polar OH1, sensors were placed on their forearm, upper arm (each 50% dominant arm) and temple (temple electrode was placed under the g.Nautilus cap and secured with a sweatband (headband) worn under the cap). Polar OH1 on the temple was placed on the same side of the body as the arm worn sensors. Centre 3 minutes of the 5 minutes resting recording were used, and the first 3 minutes of the recovery were only considered.	3-lead ECG (64-channel wireless g.Nautilus active electrode multipurpose bio signal acquisition system, g.tec medical engineering GmbH, Austria). Electrodes were attached to the participant's upper torso. Skin preparation at the electrode placement sites was performed, by cleansing with alcohol wipes and light abrasion and shaving. Silver/silver-chloride self-adhesive electrodes were placed on the participant's upper torso, under the right clavicle bone, left clavicle bone and the lower left chest regions. 1-lead ECG with sampling rate of 250Hz.	B&A, ICC
19	Horton (2017)[50]	Polar M600 Sport Watch on the left wrist. "Other Indoor training mode" or "Indoor training mode".	3-lead ECG (Power Lab 16/30 with Bio Amp model ML132) and Lab Chart Pro 7.1 Software (AD Instruments, Castle Hill, Australia). AgAgCl surface electrodes with a 19 x 16-mm active area (3M Red Dot Monitoring Electrode 2560; 3M Health Care, St. Paul, MN) were placed on cleaned skin at V2, V6, and clavicle. 1000 Hz.	B&A, mean absolute error, unbalanced repeated measures design

20	Jo (2016)[83]	Basis Peak, Fitbit Charge HR, attached to opposing wrists on the subject according to manufacturer instructions. Fitbit "track exercise" function on the mobile device application.	12-lead ECG system (Cosmed C12x; Concord, CA, USA). 10 silver/silver-chloride self-adhesive electrodes were placed on the upper torso according to the Mason-Likar-lead placement configuration.	Pearson correlation, B&A, MAPE
21	Khushhal (2017)[57]	2 Apple Watches (left and right wrists)	Polar S810i monitor	Pearson correlation, standardised mean bias, and standardised typical error of the estimate, ICC
22	Konstantinou (2020)[59]	Microsoft band 2, wrist-worn. 1 Hz sampling frequency	3-lead ECG (Biopac MP150). 2 electrodes were placed on the inner forearm of the non-dominant hand and one electrode was placed on the inner forearm of the dominant hand. Sampling frequency 1 Hz.	Pearson correlation, B&A, paired t-test, RMSE
23	Kroll (2016)[99]	Fitbit Charge HR (Fitbit, San Francisco, CA)	To provide a gold standard measurement of HR, we recovered data from the ICU bedside monitors using specialized software (BedMasterEX, Excel Medical, Jupiter, FL).	Pearson correlation, B&A, the interquartile range of differences, the median of differences
24	Menghini (2019)[36]	The E4 (Empatica) is a wrist-worn device sized 44 × 40 × 16 mm that weighs 23 g. It includes four sensors: (a) a PPG sensor that uses two green and 2 red LEDs to record blood volume pulse from the dorsal wrist (sampling frequency: 64 Hz, resolution: .9 nW/digit)	2 stainless steel (SUS03) electrodes sized 8 mm in diameter that use alternating current (8 Hz) to record skin conductance from the volar surface of the wrist (sampling frequency: 4 Hz, resolution: 1 digit ~ 900 Pico Siemens).	B&A, ICC , repeated measures mixed model
25	Müller (2019)[62]	2 wrist-worn HR trackers were used for the National Steps Challenge (Tempo HR, J-style, TEMPO) and the Polar A370 (Polar Electro Oy). Devices were worn snugly on opposite wrists (Tempo HR: left and Polar A370: right, during both the phases).	Chest-strapped Polar H10 HR monitor (Polar Electro Oy), transmitted real time HR data to a wristwatch via Bluetooth. During free-living added an ActiGraph wGT3X+BT accelerometer (ActiGraph) to collect HR data from the Polar H10 chest strap via Bluetooth.	B&A, CCC, MAPE

26	Nelson (2019)[55]	Apple Watch Series 3 (2017 version, Apple Inc, California, USA, v. 4.2.3) 42 mm was worn on the right wrist. Samples HR approx. every 10 minutes or continuously during workouts using PPG with either a green light emitting diode or infrared light and photodiode sensors. The Fitbit Charge 2 (2017 version, Fitbit Inc, California, USA. 22.55.2) was worn on the left wrist. Utilizes green LED light to continuously index HR. The Fitbit GitHub repository was used to interact with the Fitbit app programming interface to access per min data for analysis.	A standard 3-lead ambulatory ECG (Vrije Universiteit Ambulatory Monitoring System). ECG sampling frequencies were 1000 Hz, and HR was exported in 1 minute epochs, from 00 second to 59 seconds.	B&A, CCC, MAPE
27	O'Driscoll (2019)[45]	Polar m400 HR Monitor Watch and Fitbit Charge 2 (FC2) (Data are aggregated to the minute-level and synced via the Fitbit mobile application to Fitbit servers through an application programming interface. The device was fitted a finger's width above the non-dominant wrist and was configured with participant weight, height, sex and date of birth.	HR chest strap (Polar H7), transmitted second-level data via a Bluetooth connection. Data were uploaded to the Polar flow online application, then downloaded and aggregated to minute-level for analysis.	Pearson correlation, RMSE, B&A, mean absolute error, MAPE
28	Parak (2014)[91]	Mio Alpha (Mio Global, Canada), wrist-worn, data transmission ANT+ technology to Garmin Forerunner device. Schosche myRhyhm (Schosche Industries, CA, USA), forearm-worn, data transmission Bluetooth technology to iCardio Smartphone application	2-lead ECG Embla Titanium multi-parameter wearable recorder. Electrode placement: according 2 channels Holter measurement	Pearson correlation, B&A, MAPE
29	Parak (2017)[86]	Optical wrist-worn HR monitor (PulseOn, Espoo, Finland) and GPS data with a mobile phone (Samsung S3 Galaxy Trend)	Polar V800 HR monitor (Polar Electro, Kempele, Finland) with a built-in GPS sensor. Indoor: a chest strap HR device (RS800CX, Polar Electro, Kempele, Finland)	Absolute error, MAPE

30	Pasadyn (2019)[39]	Apple Watch III, FitBit Iconic, Garmin Vivosmart HR, and Tom Spark 3	3-lead ECG: The Mason-Likar electrode placement	B&A, CCC, Repeated measures mixed model analysis of variance
31	Passler (2019)[81]	2 in-ear devices: The Dash Pro (Bragi, Munich, Germany) and Cosinuss One (Cosinuss). Data was sampled at 100 Hz to the respective mobile device app.	2-lead ECG-Bodyguard 2, 1000 Hz, exported in 1 second intervals	B&A, ICC, MAPE
32	Pelizzo (2018)[100]	Fitbit Charge HR (Fitbit, San Francisco, CA, USA)	Intensive Care Unit bedside monitors (Infinity Delta, Dräger, Lübeck, Germany). Data included HR values recorded during continuous ECG monitoring, as well as HR data derived from continuous monitoring with pulse oximetry.	B&A, CCC
33	Pope (2019)[61]	Apple Watch, Fitbit Surge HR, TomTom Multisport Cardio Watch, and Microsoft Band. All wrist-worn.	Chest-mounted ActiGraph HR strap (the Polar H7 Bluetooth HR monitor; sold with the ActiGraph Bluetooth-enabled.	Pearson correlation, B&A, CCC
34	Reddy (2018)[94]	Fitbit Charge 2 and Garmin vívosmart HR+. As per the manufacturer's instructions, age, sex, height, and weight were used to initialize the wearable devices and associated applications.	Polar H7 (BTLE version) chest strap HR monitor, which was secured tightly to ensure skin contact. The data from the Polar H7 was transmitted to the Polar A300.	Pearson correlation, relative error rates, B&A, MAPE
35	Sartor (2018)[12]	Philips Electronics wrist-worn optical HR monitor	Chest strap HR monitor	B&A, mean absolute error, standard error of the estimate, bias
36	Shcherbina (2017)[46]	The Apple Watch, Basis Peak, Fitbit Surge, MicrosoftBand, Mio Alpha 2, PulseOn, and Samsung Gear S2	12-lead ECG	B&A, percent error, RMSE

37	Spierer (2015)[58]	Omron HR500U (OHR) and a Mio Alpha (MA), 2 commercial wearable HR monitors	Polar RS800CX (Polar Electro, Inc., Lake Success, NY), the chest strap was applied as per manufacturer's instructions.	Repeated-measures t-test
38	Stahl (2016)[32]	Scosche Rhythm, Mio Alpha, Fitbit Charge HR, Basis Peak, Microsoft Band, and TomTom Runner Cardio. All wrist-worn except. Scosche Rhythm (worn on the forearm with no screen readout but pairs via Bluetooth or ANT+)	Polar RS400 HR chest strap paired with a wrist receiver	Pearson correlation, B&A, MAPE, equivalence testing
39	Støve (2019)[92]	Garmin Forerunner 235	The Polar RS400, chest strap, with an inbuilt transmitter, that detects the QRS-complexes with 1 millisecond resolution and sends an electromagnetic signal to a wrist-worn watch that measures the RR interval which form the basis for the calculation of HR in bpm.	Spearman rho, Pearson correlation, ICC, B&A
40	Thomson (2019)[31]	Fitbit Charge HR 2 and the Apple Watch, placed on the left and right wrists respectively, according to the product instructions	12-lead ECG (Q-Stress ECG, Mortara, Milwaukee, WI, USA)	Relative error rates, CCC, equivalence test
41	Vandenberk (2017)[116]	The FibriCheck (Qompium) app, held against the fingertip. Converts 60 Hz video data to raw signals, which were processed with Matlab (Math-Works) to derive the corresponding PPG signal	AliveCor single-lead ECG patch attached to the upper left corner of the patient's chest with 2 disposable electrodes.	Spearman correlation, RMSE

42	Wallen (2016)[52]	Four wrist-worn devices (Apple Watch, Fitbit Charge HR, Samsung Gear S and Mio Alpha). As per manufacturer instructions, the devices were individualized for age, sex and anthropometrical data. Devices with compatible smartphone software were synchronized via Bluetooth to an appropriate smartphone to assist with data collection (ease of visualization).	3-lead ECG (CASE exercise testing system, GE Healthcare, UK). HR from the ECG and devices was manually recorded every 15 seconds during the protocol	Spearman correlation, Pearson correlation, B&A, ICC
43	Wang (2016)[117]	Wrist-worn: the Mio Alpha (Mio Alpha; Physical Enterprises Inc., Vancouver, BC). Two green LED lights that shine into the skin, and an electro-optical cell senses the changes in the colour of the skin, i.e., the blood flow. Algorithms are applied to the blood flow signal and HR is derived. When measuring HR, the watch simultaneously transmits the measured data to smartphones or laptop over Bluetooth 4.0. A computer program was developed to receive real time HR data from Mio Alpha and store the data with local timestamps.	Physiological status monitor chest strap (Bioharness (version 1); Zephyr Technology Corp., Annapolis, MD). The Bioharness system uses a single-channel ECG sensor and circuitry to measure HR through RR interval calculations at a sampling rate of 250 Hz. Measured data are offline recorded in the module memory (512M,~480 hr).	Pearson correlation, B&A, standard error of the estimate
44	Zheng (2012)[82]	3 PPG devices, placed on right ear lobe, right index finger and nose bridge (eyeglasses-based). The pass band of the analogue band-pass filters applied on the PPG signals from 0.5 to 15 Hz.	ECG. The pass band of the analogue band-pass filters applied on the PPG signal is from 0.5 to 15 Hz.	Student's t-test

Abbreviations. HR: heart rate; ICC: intra class correlation coefficient; B&A: Bland-Altman analysis; MAPE: mean absolute percentage error; RMSE: root-mean-square error; ECG: electrocardiogram; LED: light-emitting diode; Hz: Hertz; CCC: Lin's concordance correlation coefficient; PPG: photoplethysmography; bpm: beats per minute.

Table 6. Summary of data handling methodologies.

N	Author (year)	Smoothing of index test data	Smoothing of criterion measure data	Motion artefacts	Data synchronization	Excluded data
1	Abt (2018)[88]	HR data was recorded every 5 seconds.	Criterion HR was measured using a Polar T31™ chest strap interfaced with a metabolic cart.	ND	The "Workout" app automatically syncs exercise data to the "Health" database on its paired iPhone after the completion of an exercise session.	Missing HR data was excluded on one occasion as the Polar T31™ monitor did not record HR.
2	Bai (2018)[93]	HR data from the Fitbit Charge HR was accessed through the third-party website Fitabase (Small Steps Labs LLC., San Diego, CA).	Minute by minute	ND	ND	ND
3	Boudreaux (2018)[95]	ND	ND	ND	Readings from all wearable devices were digitally time stamped to an iPhone 7 Plus in the Apple Health application and/or in the device's specific application. HR was recorded from the ECG at each time point and confirmed by measuring the distance between R and R waves in consecutive cadence cycles from hardcopy ECG printouts.	ND
4	Brazendale (2019)[60]	Data from the Fitbit Charge HR© was downloaded via a third-party research platform, Fitabase©.	Data downloaded via manufacturer software.	ND	Prior to data collection, the time for the Fitbit Charge HR© and the Polar H7© watch was calibrated to the nearest second.	Data was cleaned for the removal of corrupt files due to criterion measure device malfunction.
5	Cadmus-Bertram (2017)[89]	ND	ND	ND	ND	ND
6	Claes (2017)[53]	ND	ND	ND	The Garmin Forerunner 225 was started simultaneously	ND

					with the start of the test. This time point was also manually written down by a second researcher to allow identification of the start point of the test in the Zensor data. Raw HR data was obtained offline through the Zensor software.	
7	Coca (2010)[38]	ND	ND	ND	Physiological data are stored onto a small, portable data recorder carried in a pouch attached to the vest, and telemetered in real-time to a laptop computer.	ND
8	Dooley (2017)[40]	ND	ND	ND	ND	ND
9	Dur (2018)[37]	ND	No digital filtering was applied to the raw ECG.	Segments of the PPG signal containing artefacts related to wrist movement were removed.	The synchronous recordings from ECG and Wavelet wristband devices were aligned manually based on time stamps and agreement of interbeat intervals, although a small misalignment was inevitable because of the lacking information on the pulse transit time.	For several participants (n=12), the test was halted before the 3 minute mark because of discomfort while breathing into the spirometer.
10	Etiwy (2019)[33]	Of the 2,560 possible HR measurements (80 participants, 8 time points, 4 devices per subject (ECG, Polar chest strap, two wrist-worn monitors)), 2,546 were recorded (99.5%). Missing data were attributable to failure of the device to display/record HR (5 for	ECG-based HR was determined by visual assessment under direct supervision by a cardiologist; ECG-based HR was able to be ascertained at all time points, and ECG artefact was not observed.	ND	ND	ND

		Apple Watch and 9 for TomTom Spark Cardio).				
11	Falter (2019)[85]	ND	ND	ND	ND	ND
12	Georgiou (2019)[34]	ND	ND	ND	All the time points had to be converted to absolute local time. Additionally, since both devices did not share the same time settings from a reliable third-party source, their recorded data needed synchronization.	ND
13	Gillinov (2017)[41]	Processed according to proprietary algorithms	ND	Across all ECG tracings, there was minimal artefact and in no situation did ECG artefact interfere with visual HR determination.	ND	Missing data were attributable to failure of the device to record HR (8 for Apple Watch, 4 for Fitbit, two for Scosche Rhythm+, and one for Garmin Forerunner 235
14	Gorny (2017)[98]	Fitbit HR measures were downloaded directly from the Web server using a developer's application programming interface issued by Fitbit.	To record the Polar H6 HR (Polar) data, these participants were provided with an Actigraph GT3X+ logger on Bluetooth receiver mode set to sample measures at 10 second intervals.	ND	ND	All 1 minute epochs measuring non-zero HR were included.
15	Hahnen (2020)[35]	ND	ND	ND	ND	Excluding data from 42 individuals because of excessive variation in sequential standard measurements per prespecified dropping rules. Excluded data from participants with a variation in standard measurements

						greater than 12 mm Hg for systolic blood pressure and 8 mm Hg for diastolic blood pressure, in accordance with validation guidelines.
16	Hendrikx (2017)[54]	All data were resampled to a common 1 Hz resolution	All data were resampled to a common 1 Hz resolution	ND	The 1 minute average values for HR, and cumulative steps and energy expenditure over 1 minute, are logged in internal memory and transmitted via Bluetooth to a phone running the companion app for 24/7 monitoring.	2 participants were excluded due to a history of epilepsy. 2 participants experienced an adverse event that was classified as non-serious and not device-related after assessment by the trial's independent medical monitor. Some data of participants were excluded from specific analyses because data were not correctly logged or, based on objective criteria, were found to be invalid.
17	Hermand (2019)[56]	Smoothed on a 10 second window.	Smoothed on a 10 second window.	ND	Recordings for both were started at rest before the exercise start and terminated after a short recovery time. signals were synchronized with the least square method.	Visually inspected for criterion dysfunction, discarded when necessary.
18	Hettiarachchi (2019)[90]	A custom data logger was developed to interface simultaneously to the 3 Polar OH1 sensors utilizing Bluetooth Low Energy technology. The logger software exported the time stamped HR measurements of the 3 Polar sensors to a CSV comma separated file for off-line processing.	0.1 - 100 Hz bandpass filter and a 50 Hz notch filter. ECG recordings with extremely noisy signals were manually marked and excluded. Subsequently, the QRS complexes of the ECG signals were detected using the Pan-Tompkins QRS detection algorithm. Then the R-peak series (tachogram) was obtained by calculating the intervals between successive R-peaks (RR interval). The R-	ND	ND	At some instances, the Polar OH1 data measurements were missing due to low skin contact or loss in Bluetooth connection. On average about 5% of the data was lost from the Polar measurements.

			peak series is then examined and corrected for any missed and/or extra beats using a quotient filter.			
19	Horton (2017)[50]	HR data was downloaded at 1 second intervals using the Polar Flow Web service.	ECG data were sampled at 1000 Hz to display the PQRST waveform in Lab Chart Pro 7.1. Using an algorithm in Lab Chart Pro 7.1, HR was calculated from the time between the RR intervals. ECG HR data were then down-sampled from 1000 Hz and exported as a text file at 1 second intervals.	ND	"A ‘‘start’’ marker was inserted in Lab Chart Pro 7.1 to be used later for synchronization of ECG and Polar M600 HR data. The two HR data files for each subject were synchronized by using the start marker in the ECG data file and the first Polar M600 HR sample. Every 10 seconds throughout the data files, mean HR was calculated for both measurement devices."	ND
20	Jo (2016)[83]	ND	HR data per second was converted to bpm automatically by the data acquisition software program prior to analysis.	ND	"Time synced HR data from each device (test devices and ECG) were concurrently and continuously acquired second by second throughout the entire 77 minutes protocol for each participant. Data acquisition from each device along with ECG was time-synced according to a single master clock."	Initial rest period
21	Khushhal (2017)[57]	The ‘Workout’ app nominally records HR at 5 second intervals. On cessation of each trial the HR data were synced automatically to the ‘Health’ database on its paired iPhone.	The sampling time for the Polar S810i HR monitor was set at 5 second intervals. Following exercise, the HR data were transferred from the Polar S810i HR monitor to the Polar Pro Trainer 5 software.	ND	ND	ND
22	Konstantinou (2020)[59]	ND	1) manually: raw ECG signals were filtered by a BIOPAC	ND	Stationary data were analysed traditionally in	ND

			ECG100C bioamplifier, which was set to record HR from 40 to 180 bpm. 2) automatically: HR data was conducted in AcqKnowledge based on its internal algorithm (name not reported by developers).	AcqKnowledge based on its internal algorithm (name not reported by developers). Mean values were extracted into Excel. For the automated analysis, the Acq files of the raw stationary data were read by our Python program, and their mean values were computed. For the wearable device, raw data were computed internally by the Microsoft band 2, and then, their mean values were computed using the same Python program as in the stationary automated analysis. The mean HR were calculated for an interval of every 10 second for each of the phases, for both the wearable and stationary devices.	
23	Kroll (2016)[99]	Automated Python script to derive minute-level HR	ND	ND	Synchronized bedside monitor data and personal fitness tracker data using a correction factor that accounted for the difference between each device's internal clock.	2 patients whose devices were removed early.

24	Menghini (2019)[36]	Automatically detection "find peaks" (manually corrected) automatic detection and removal of artefacts (algorithm: Berntson et al., 1990) and further visual correction	Automatically band-pass filtered (0.05 Hz–1 kHz) down-sampled to 256 and 4 Hz	ND	Synchronization between recordings was performed by marking 3 events in both systems simultaneously, prior to each session. The average time difference between the two systems was added to the Infiniti scripted time stamps to obtain the corresponding condition-related epochs in the E4 data. Synchronization was verified by visual comparison of acceleration time trends, and data with a considerable time shift were discarded.	Data with considerable time shift in the synchronization phase was discarded. Low-quality signal (skin conductance) was discarded. 10 participants were excluded for different reasons: ectopic beats in more than 50% of the recording (N = 2), wristband troubleshooting (N = 1), technical problems in the standard recording system (N = 4), or failed synchronization between the two systems (N = 3).
25	Müller (2019)[62]	The sampling frequencies of the Tempo HR, Polar A370, and Polar H10 chest strap were 0.1 Hz, 1 Hz, and 1 Hz, respectively. As such, HR data was collected every second by the Polar devices and every 10 seconds by the Tempo HR.	The sampling frequencies of the Tempo HR, Polar A370, and Polar H10 chest strap were 0.1 Hz, 1 Hz, and 1 Hz, respectively. As such, HR data were collected every second by the Polar devices and every 10 seconds by the Tempo HR.	ND	All devices provided time-stamped HR data based on the Network Time Protocol (GMT plus 8 hours). This allowed for time matching of data.	Due to the unavailability of some HR data, few participants were excluded from some analyses.
26	Nelson (2019)[55]	During workout, the average HR per minute was used.	Averaged in 1 minute intervals	ND	ND	Outliers were not removed as this would interfere with determining device accuracy during consumer use conditions.
27	O'Driscoll (2019)[45]	Data are aggregated to the minute-level and synced via the Fitbit mobile application to Fitbit servers through an	Data was uploaded to the Polar flow online application, then downloaded and aggregated to minute-level for analysis.	ND	ND	ND

		application programming interface.				
28	Parak (2014)[91]	Signals were smoothed by moving average in 5 second window.	"Analysis by Kubios HRV tool. The better ECG raw signal quality channel was selected by visual inspection of both recorded channels. The R-peaks were detected in selected channel by automatic R-peak detection algorithm which is included in HRV tool. Signals were smoothed by moving average in 5 second window."	ND	The evaluated and reference HR signals were resampled to 10 Hz sampling frequency. HR acquired from PPG HR monitors and reference HR were synchronized in time by applying cross-correlation function between the reference and the target HR and by maximizing the cross-correlation value at t=0.	Heart timing signals algorithm was used for detection of the arrhythmias (ectopic beats). These beats were excluded from the final statistical evaluation and error estimation.
29	Parak (2017)[86]	ND	After applying an artefact correction algorithm to the signals, the maximum HR value was observed.	ND	HR signals were synchronized in time by maximizing the cross-correlation between the signals at t=0.	ND
30	Pasadyn (2019)[39]	Proprietary algorithm to determine changes in blood volume based upon reflected light.	ND	ND	ND	ND
31	Passler (2019)[81]	ND	Integrated algorithm to correct artefacts.	Motion artefacts, due to the change of body position and the re-adjustment of the sensors, resulted in strong signal noise. Consequently, this data was not considered in the statistical evaluation.	Data files of the in-ear and ECG devices were synched using the respective timestamps of each data acquisition. All data files were recorded in the Unix timestamp format (UTC). This format counts time in millisecond since 1 January 1970. In contrast, the Dash Pro counts time in millisecond since 1 January 2015. This represents an overall time discrepancy of 45 years or a shift by up to 40 seconds within 24 h. This	Motion artefacts, due to the change of body position and the re-adjustment of the sensors, resulting in strong signal noise.

					correction was carried out immediately before each examination.	
32	Pelizzo (2018)[100]	ND	ND	ND	We synchronized the bedside monitor data and PFT.	ND
33	Pope (2019)[61]	Two researchers collected these smartwatch HR and EE data from the smartwatches immediately after each participant finished their respective exercise session—allowing each participant's smartwatch data from each exercise session to be double-checked (i.e., data quality control protocol).	HR analysis was completed concurrently using a 1-second epoch, with HR data exported from ActiLife to a Microsoft Excel Spreadsheet for average/peak HR calculation, with all HR data trimmed to include only the 20 minutes exercise session and reviewed for physiologically implausible values.	ND	Given that these data were collected directly from the smartwatches, no syncing issues were encountered.	ND
34	Reddy (2018)[94]	Garmin: According to the device specifications, the frequency at which HR is measured is normally once every 15 seconds, but triggering the device key button and setting the wearable to an activity mode (e.g., run) increases the frequency at which HR is measured. Fitbit: According to the manufacturer, the frequency at which HR is measured during activity mode is once every second. Data were	ND	ND	Synchronization of all the devices to a single clock before the exercise protocol commenced.	ND

		downloaded at the highest sample rate possible through Fitabase (Small Steps Labs, California, US), a third-party research platform designed to collect data from Fitbit using the developer application programming interface.				
35	Sartor (2018)	Optical HR monitor logged the PPG data (16, 32, 64 or 128 Hz). Real-time HR computation was based on a 5 second sliding window. Estimated HR and a HR quality index were logged together every second. The data were stored in the internal memory of the prototype. These data were transferred via USB onto a personal computer at the end of each test.	Radio connected to a logging watch. The chest strap was set to output a HR every 5 seconds.	Highly periodic activities showed a higher data coverage than less periodic activities. Highest data coverage was found in activities with the lowest effect of motion artefacts (cycling, sedentary).	In the automated process the two sequences were interpolated on a uniform time grid by linear interpolation. The delay was calculated as the location of the maximum of the cross covariance function between the interpolated sequences, and the sequences were then aligned. A final visual inspection was performed to check the alignment and to discard erroneous reference data.	Data coverage did not fall below 92.2%
36	Shcherbina (2017)[46]	Principal component analysis to identify outliers and cluster errors. Singular value decomposition over the activity error rates. 3 regression approaches were applied to uncover associations in the dataset.	ND	ND	ND	Participants with missing data were excluded from the principal component analysis.

37	Spierer (2015)[58]	The Polar, Omron and Mio Alpha devices collected data in 5 second epochs, which were used to calculate the average HR over each minute while performing the study tasks. Noise removal algorithm.	ND	The signal processing algorithm measures HR continuously during exercise by removing the motion artefact.	Based on 5 second intervals of data collection, values from the Polar, HR500U and Mio Alpha were synchronized to directly compare data from all devices.	ND
38	Stahl (2016)[32]	ND	ND	ND	ND	ND
39	Støve (2019)[92]	ND	ND	ND	HR was concurrently assessed with both monitors and manually recorded by a researcher taking a digital picture every 60 seconds with both monitors' in the same frame thus ensuring that criterion measures were obtained simultaneously.	Simultaneous HR measurements were made every minute and data from the last measurement in each activity level was used for analysis.
40	Thomson (2019)[31]	ND	ND	ND	HR readings were taken manually from each device and from the ECG each minute for the entire duration of the exercise protocol.	ND
41	Vandenberk (2017)[116]	ND	ND	ND	Time synchronization between ECG and PPG was automatically done by the FibriCheck app	A total of 3 persons were excluded from analysis because of failure to obtain valid data with 1 or more devices.
42	Wallen (2016)[52]	ND	ND	ND	ND	All participants wore each device once however EE data were missing for 3 participants and step count data were missing for two due to a data recording error.
43	Wang (2016)[117]	In order to reduce the effect of noise, each	In order to reduce the effect of noise, each minute was	ND	A laptop (Intel Core i7 CPU @ 2.8GHz, 4GB	ND

		minute was divided into 6 10 second intervals and mean HR over the third and sixth intervals were calculated for comparison between devices.	divided into 6 10 second intervals and mean HR over the third and sixth intervals were calculated for comparison between devices.		RAM,500GB HDD, Bluetooth 4.0) was used to receive real-time HR data from Mio Alpha and synchronize the Internal clock of Bioharness system. It provides a unified time reference for data measured by the 2 devices.	
44	Zheng (2012)[82]	The acquired ECG and all PPGs were filtered by low-pass filter with cut off frequency at 30 Hz and 16 Hz, respectively.	The acquired ECG and all PPGs were filtered by low-pass filter with cut off frequency at 30 Hz and 16 Hz, respectively.	Distorted PPG waveform due to motion artefacts was manually removed and the corresponding HR and pulse transit time values were excluded from the analysis.	ND	Distorted PPG waveform due to motion artefacts was manually removed and the corresponding HR and pulse transit time values were excluded from the analysis.

Abbreviations. HR: heart rate; ND: not disclosed; ECG: electrocardiogram; PPG: photoplethysmography; Hz: Hertz; bpm: beats per minute; HRV: heart rate variability; EE: energy expenditure.

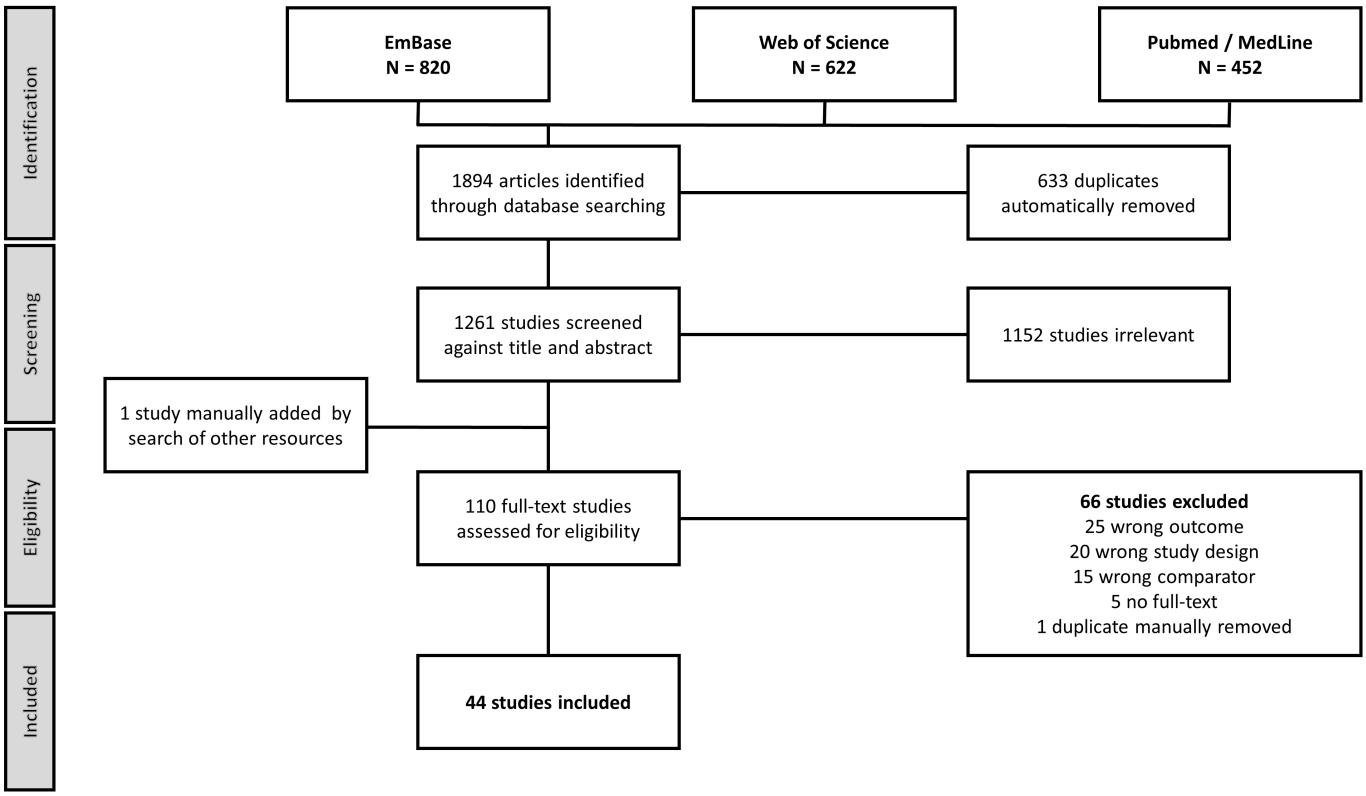
Table 7. Examples of validated chest strap devices for the measuring of RR intervals.

Author (year)	Index test	Criterion measure	Participants	Activity protocol	Statistics	Validity
Chellakumar (2005)[118]	Polar T31 (Polar Electro Oy, Kempele, Finland)	A 3-lead system (BIOPAC Systems Inc., CA). 1000 Hz	7 healthy male subjects (age = 23.5 (mean) ± 4.5 (SD) years; height = 1.77 ± 0.1 m; weight = 74.7 ± 10.7 kg)	Acclimated in a dark, ambient environment for 15 minutes. Sit and remain stationary for 15 minutes. Sound-attenuating headphones were worn to minimize interference from the external environment	ANOVA	Was found to be comparable to ECG for HRV measurements

Engström (2012)[119]	Polar RS400 (Polar Electro Oy, Kempele, Finland)	ECG (CS-200 Ergospirometry, Schiller AG, Altgassse 68, CH-6341 Baar Switzerland) using standard 12-lead, was measured with 6 electrodes	10 healthy subjects, 19 - 34 years	The exercise test was performed on a cycle ergometer (Monark 839E). Subjects cycled for 5 minutes at each of three power levels, 50 W, 100 W and 150 W, with no rest in between	Pearson correlation, student's paired t-test, B&A. repeatability coefficients	Significant linear relationships, correlation coefficients between 0.97-1.0. T-tests revealed no differences. Mean difference ± 2SD between the methods was 0.7 ± 4.3 bpm in test 1 and 0.2 ± 3.2 bpm in test 2
Gilgen-Ammann (2019)[120]	Polar H10 HR monitor with a Pro Strap (Polar Electro Oy, Kempele, Finland)	Schiller medilog® AR12plus ambulatory 3-lead ECG Holter monitor (Schiller Medizintechnik GmbH, Baar, Switzerland). 1000 Hz	10 (5 females and 5 males) healthy, lean, and physically fit volunteers (age 24.7 ± 1.9 years, body height 172.5 ± 8.4 cm, body weight 67.5 ± 9.7 kg, BMI 22.6 ± 1.3 kg/m², and chest circumference 80.3 ± 6.8 cm)	(1) sitting in a chair and reading (sedentary activity); (2) wiping the floor with a mop and hanging out the laundry at a self-guided order and pace (household chores); (3) normal walking on a treadmill at 5.5 km/h; (4) jogging on a treadmill at 11 km/h; and (5) a strength training circuit of 5 aligned 60 second cycles with 45 second workouts and 15 seconds rests, including squats, shoulder shrugs, bicep curls with a dumbbell in each hand (4.5 ± 1.6 kg), lunges, and sit-ups	Spearman correlation, Wilcoxon test, B&A	In terms of the measurement agreement, a high correlation was found (r=0.997), and in 97.1% of the measured RR intervals, the values provided by both systems differed less than 2% among each other
Romagnoli (2013)[121]	The GOW system (Weartech sl., Spain)	1000 Hz. Cardiolab II plus (ECG) (Prucka engineering, TX, USA)	12 adult male volunteers aged between 52 and 66 years [age 60.8 ± 5.76 years, height 174.2 ± 7.1 cm, weight 65.6 ± 6.5 kg, BMI 21.6 ± 1.5 kg/m², mean ± standard deviation (SD)]. They all suffered acute myocardial infarction	1 minute warm-up of 45 W pedalling followed by a gradual increase of load until each patient's target HR was reached. During the test the cadence was set free between 60 and 90 rpm. Target HR set by medical staff based on patient's pathology	B&A, ICC and 95% CI	Limits of agreement (LoA) on RR intervals were stable at around 3 milliseconds. The GOW system is a valid tool for controlling HR during physical activity (not HRV)

Weippert (2010)[71]	Polar S810i (Polar Electro Oy, Kempele, Finland) and Suunto t6 (Suunto Oy, Finland, 1000 Hz)	Ambulatory 5-lead 2-channel ECG system (Cardiolight S, Fa. Medset, Germany) providing a sampling rate of 200 Hz and a temporal resolution of 5 milliseconds	19 young males (aged between 22 and 31 years, median 24 years)	10 minutes in supine rest, 10 minutes of light dynamic exercise (walking) and 5 minutes of moderate to vigorous isometric muscular exercise of the upper and lower limb six times with 5 minutes of sitting rest between each exercise	ICC and 95% CI, B&A	Regarding the RR interval recordings, ICC (lower ICC 95% CI >0.99) as well as LoA (maximum LoA: -15.1 to 14.3 milliseconds for ECG vs. Polar) showed an excellent agreement between all devices

Abbreviations. HZ: Hertz; SD: standard deviation; ECG: electrocardiogram; HRV: heart rate variability; W: Watt; B&A: Bland-Altman analysis; bpm: beats per minute; HR: heart rate; BMI: body mass index; rpm: repetitions per minute; ICC: intra class correlation coefficient; CI: confidence intervals; LoA: limits of agreement.


Identification
Screening
Eligibility
Included
EmBase
N = 820
Web of Science
N = 622
Pubmed / MedLine
N = 452
1894 articles identified through database searching
633 duplicates automatically removed
1261 studies screened against title and abstract
1152 studies irrelevant
1 study manually added by search of other resources
110 full-text studies assessed for eligibility
66 studies excluded
25 wrong outcome
20 wrong study design
15 wrong comparator
5 no full-text
1 duplicate manually removed
44 studies included