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

Mühlen JM, *et al. Br J Sports Med* 2020;0:1–13. doi: 10.1136/bjsports-2020-103148

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

Mühlen JM, *et al. Br J Sports Med* 2020;0:1–13. doi: 10.1136/bjsports-2020-103148

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)

Mühlen JM, *et al. Br J Sports Med* 2020;0:1–13. doi: 10.1136/bjsports-2020-103148

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)

Mühlen JM, *et al. Br J Sports Med* 2020;0:1–13. doi: 10.1136/bjsports-2020-103148

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

Mühlen JM, *et al. Br J Sports Med* 2020;0:1–13. doi: 10.1136/bjsports-2020-103148

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

Mühlen JM, *et al. Br J Sports Med* 2020;0:1–13. doi: 10.1136/bjsports-2020-103148

		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.