Original research article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 12: 1–17

DOI: 10.1177/20552076261426622
journals.sagepub.com/home/dhj

Sage

A novel machine learning procedure to detect and remove artefacts in heart rate data obtained from photoplethysmography wearables: A prospective cohort study

Paulien Vermunicht[1,2], Christophe Buyck[1,2], Sebastiaan Naessens[2], Wendy Hens[2,3], Emeline Van Craenenbroeck[1,2], Juan Sebastian Piedrahita Giraldo[4,5], Katsiaryna Makayed[4,5], Saartje Herman[4,5], Kris Laukens[4,5], Johan Roeykens[6], Koen De Deckere[7], Lien Desteghe[1,2,8,9] and Hein Heidbuchel[1,2,8]

Abstract

Introduction: Photoplethysmography (PPG) based heart rate (HR) monitoring supports continuous assessment of physical activity intensity, but device generated HR values remain prone to motion artefacts. This study validated a novel machine learning procedure that detects artefacts directly in PPG derived HR data, rather than in raw waveforms that are unavailable in commercial devices.

Methods: In this prospective study, 149 participants (46 following cardiac rehabilitation, 103 healthy) wore a PPG-based wrist device and a reference chest strap for 12 weeks. Prior testing defined participants as "PPG-compatible" (i.e., HR error <10% during ≥70% of training data). Three quarters trained artefact and activity detection models, one quarter served as the test population. To balance artefact removal and activity preservation, multiple probability thresholds were combined to reject unreliable HR values, remove or interpolate using adjacent reliable values. The Antwerp Activity Index (AAI), an HR-based PA score, was calculated from the resulting PPG and reference data to evaluate threshold combinations.

Results: Seventy-eight participants (53.8%) were PPG-compatible, yielding over 5 million datapoints, with 75% ($n = 58$; 4,144,654 datapoints) used for training and 25% ($n = 20$; 992,180 datapoints) for testing. The artefact model detected artefacts with a sensitivity of 58.5% in daily life and 76.6% during exercise, while limiting incorrect removals (74.9–91.3% specificity). Rejecting data for AAI-calculation when artefact probability was >50% and activity probability <70% produced the highest agreement with reference AAI (Pearson's $r = .89$–$.96$).

Conclusions: This new machine learning-based procedure, which operates on device-generated HR rather than PPG waveforms, effectively removes artefacts while preserving key activity signals, ensuring clinically relevant PA assessment in daily life.

[1]Research Group Cardiovascular Diseases, University of Antwerp, Antwerp, Belgium
[2]Department of Cardiology, Antwerp University Hospital, Antwerp, Belgium
[3]Research Group Movant, University of Antwerp, Antwerp, Belgium
[4]Department of Computer Science, University of Antwerp, Antwerp, Belgium
[5]Biomedical Informatics Research Network Antwerp (Biomina), University of Antwerp, Antwerp, Belgium
[6]Department S.P.O.R.T.S., Antwerp University Hospital, Antwerp, Belgium
[7]Sports Medical Center Nottebohm Brecht, Brecht, Belgium
[8]Faculty of Medicine and Life Sciences, Hasselt University, Hasselt, Belgium
[9]Centre for Research and Innovation in Care, University of Antwerp, Antwerp, Belgium

Corresponding author:
Paulien Vermunicht, Department of Cardiology, Antwerp University Hospital, Drie Eikenstraat 655, 2650 Edegem, Belgium.
Email: Paulien.Vermunicht@uantwerpen.be

Keywords
Heart rate, photoplethysmography, artefact detection, wearable devices, cardiac rehabilitation, exercise

Received: 5 September 2025; accepted: 28 January 2026

Introduction

Accurate heart rate (HR) monitoring is a valuable tool for objectively tracking physical activity (PA) intensity, a key factor in both primary and secondary prevention of cardiovascular diseases (CVD).[1–3] HR monitoring is widely integrated into structured cardiac rehabilitation (CR) programmes to guide patient-specific exercise prescriptions[4] and is also frequently used in healthy populations to optimise training intensity and track fitness progress.[5]

Wrist-worn devices using photoplethysmography (PPG) technology offer an accessible solution for near-continuous HR monitoring, but can be prone to motion artefacts. These artefacts can result in over- or underestimation of true HR, which may lead to incorrect PA intensity estimation and limit reliability and clinical utility for PA monitoring.[5,6] Distinguishing artefacts from true physiological HR changes remains a challenge in both daily-life and exercise conditions, where uncontrolled movement introduces noise into the data.[7,8]

Machine learning (ML) approaches are increasingly applied to enhance data quality and clinical interpretability, and offer promising solutions for improving the reliability of PPG-derived HR data by identifying and removing artefacts.[9] While previous research has primarily focused on removing artefacts from raw PPG waveform signals, these algorithms are often proprietary and possibly patented, either embedded in commercial devices or available only through specialised research equipment.[10–13] In contrast, our research group was the first to explore ML-based artefact detection directly at the level of device generated HR values from commercially available wrist worn devices, where raw PPG waveforms are not accessible, aiming to enhance the usability of PPG devices for PA assessment in clinical settings. We developed a novel ML-based artefact removal procedure (ARP) designed to differentiate artefactual HR fluctuations from true PA-related HR variations and to eliminate unreliable data points. A pilot study demonstrated the potential of this approach (e.g., artefact detection accuracy: 80% based on cross-validated training data); however, its validation was limited by a small sample size, no external testing dataset, and few diverse activity intensities.[14–16]

To address these limitations, the present study aimed to optimise and validate the ARP in a larger and more physically active cohort. By training independent ML models for artefact detection and activity recognition on a diverse dataset, our goal was to balance the removal of erroneous HR data with the preservation of meaningful, activity-related data. The primary objective was to validate the ARP across daily life and exercise conditions using standard classification performance metrics. In addition, we evaluated the impact of the ARP on an HR-based PA scoring metric, the Antwerp Activity Index (AAI), to assess its clinical utility. By integrating ML-based artefact detection with PA quantification, we aimed to enhance the reliability of PPG-based HR data for assessing PA in CR settings and the daily life of ambulatory cardiac patients.

Methods

A prospective cohort study (ARTEPHYSICAL study, NCT05901038) was conducted to train and validate an ML based ARP to detect artefacts, recognise activity episodes, and label unreliable data in continuous HR data obtained from a PPG based wearable device. The study was approved by the Ethics Committee of University Hospital Antwerp (UZA) and the University of Antwerp, and conducted in accordance with the Declaration of Helsinki (EC reference: 2023-5241, BUN: B3002023000046).

Study population

A total of 149 participants were included in the study, divided into three groups: (1) patients enrolled in a CR programme after myocardial infarction, percutaneous coronary intervention (PCI), cardiac ablation, or cardiac surgery ($n = 46$); (2) coached healthy individuals who received a structured 12-week home-based training schedule developed by Sport Medical Centre Nottebohm ($n = 57$); and (3) non-coached healthy individuals attending a routine sports medical check-up, but without entering a structured training programme ($n = 46$).

A formal sample size calculation was performed during the design of the broader ARTEPHYSICAL study (NCT05901038), based on a separate planned analysis examining the correlation between the Antwerp Activity Index (AAI, discussed later) and maximum oxygen uptake (VO_2max). Although that analysis falls outside the scope of the current manuscript, it informed our target of 46 participants per group. The slightly larger number in the coached healthy group ($n = 57$) reflects high recruitment interest and was intended to compensate for a higher observed dropout in that subgroup. The resulting dataset comprises over 5 million time-matched HR datapoints for ARP training and testing (see later), providing sufficient scale and diversity for robust algorithm development.

Participants were eligible if they were ≥18 years old, possessed a smartphone, and provided written informed consent. Exclusion criteria included severe heart failure (New York Heart Association classification III–IV), inability to understand Dutch or English, or cognitive impairment (e.g., severe dementia). These three groups were specifically chosen since they engage in activities in their rehabilitation trajectory, training schedule, or leisure time. This ensured that the ARP could be trained and validated on a diverse dataset encompassing both rest and various types of PA, thereby optimising its capacity to distinguish artefacts from true physiological HR fluctuations.

Study procedure & devices

Participants were monitored for an average period of 12 weeks using a PPG-based wrist-worn device (Fitbit Inspire 2, Fitbit, Inc., San Francisco, CA, USA) and a reference chest strap (Polar H10, Polar Electro Oy, Kempele, Finland). The Fitbit device was selected because it is one of the few commercially available PPG wearables that provide access to continuous HR data with sufficient temporal resolution via an API service. Other commonly used devices typically provide only aggregated HR summaries, which would not allow the detailed HR based analyses required for this study. The reference device, the Polar H10 chest strap, measures HR using a single-lead electrocardiographic (ECG) signal via two skin-contact electrodes.[17] The H10 has been shown to provide highly accurate RR-interval and HR measurements and demonstrates excellent agreement with Holter ECG recordings in both healthy individuals and patients with cardiac disease.[4,14,16,18,19] This chest strap was preferred over a traditional Holter ECG, which is less practical for repeated or prolonged use during a 12-week protocol.

The study team assisted in installing both HR monitors and corresponding smartphone applications (Fitbit & Polar Beat) and provided standardised instructions. Participants were advised to wear the PPG-based wrist-worn device three finger widths above the wrist joint, ensuring that it was tight but comfortable, based on previous research.[15] The PPG device measured HR passively in the background, continuously throughout the 12-week study period. The reference chest strap was worn continuously for one 24-h monitoring period, representing daily life, and subsequently only during exercise. For each session, participants were instructed to manually start and stop the recording in the Polar Beat app. Exercise was defined as any PA lasting ≥10 min, including but not limited to CR training sessions, walking, cycling, fitness, gardening, and household activities. Participants also recorded their daily activities in a diary during the first study week, with an optional extension if they were willing to.

For both devices, data synchronisation to pseudonymised accounts occurred automatically via Bluetooth and Wi-Fi. To minimise data loss, the study team monitored synchronisation status weekly, and participants were asked to manually synchronise via the Fitbit and Polar smartphone applications if needed (i.e., by swiping down on the application's home screen).

Data collection and preprocessing

HR data from both devices were exported as CSV files: Polar data via the Polar Flow web service and Fitbit data via the Fitbit Web API. Sampling frequencies were 1-s intervals for Polar and 5-s intervals for Fitbit. Short deviations from the 5-s sampling grid occurred in the exported Fitbit data and were addressed with interpolation as described below. Extreme outliers beyond the range of physiological plausibility ($HR \leq 25$ b/min, $HR \geq 220$ b/min) were removed from the data. Next, missing HR values were estimated using polynomial interpolation for gaps <30 s, while longer gaps remained unfilled. Across all participants and over their full monitoring period, 38.9% of all values in the final Fitbit HR time series were imputed. Importantly, interpolation was almost exclusively driven by very short gaps: 69.3% of imputed values originated from 15-s gaps (two consecutive missing 5-s samples) and 29.9% from 10-s gaps (one missing sample). Longer gaps were rare; gaps of 30 s or more accounted for only 0.2% of all identified gaps and were not imputed. The proportion of interpolated samples per participant showed a median of 38.5% (interquartile range 36.4–40.9%), illustrated in Supplemental Figure 1.

PPG-compatibility analysis

Our previous research demonstrated that PPG-based HR monitoring is not feasible for all individuals, as some exhibit persistently poor accuracy despite technical optimisations and artefact removal.[15,16] Since excessive noise could hinder model training,[20] a PPG-compatibility analysis was performed for each participant and the ML models were trained exclusively on PPG-compatible individuals.

PPG-compatibility was assessed based on HR data from the first three training sessions lasting >20 min. The selection of these sessions was also informed by the diary-reported activity content and varied across groups: for CR patients, supervised CR training sessions were prioritised, while for the coached and non-coached participants, structured sports-related activities (e.g., running and cycling) were preferred over non-sport activities (e.g., household chores and working in the garden). If fewer than three training sessions were available, the analysis was performed on the available data. Participants were classified as PPG-compatible if at least 70% of their training HR data was accurate (defined as mean absolute percentage error, MAPE <10%). Those with <70% accurate training data were classified as PPG-incompatible. The 10%

MAPE threshold was based on previous research and HR monitoring standards,[21–23] while our clinical team selected the 70% threshold as a practical benchmark for reliable HR monitoring during exercise.

Training and optimisation of the ARP: artefact and activity models

The ARP was designed to distinguish artefactual fluctuations in PPG derived HR values from PA related HR changes and to remove unreliable data. A first version of this ML-based procedure was described in detail in a previous pilot study.[16] In the current study, we substantially optimised the ARP and trained it on a larger, more heterogeneous dataset to improve its generalisability across activity types and individuals.

In short, the ARP consists of two independent ML classification models, one for artefact detection and one for activity recognition. Their outputs are subsequently combined to identify unreliable data points (see section Combining prediction scores to label unreliable data and achieve clinically relevant HR data). Supplemental Table 1 provides a full overview of all modifications made to the models' characteristics compared to the previous version. Below, we summarise the key characteristics and updates of the current models.

Target artefact and activity labels, necessary to guide the training process and evaluation, were assigned using reference device data and activity diaries. Datapoints were labelled as true/target artefacts when the mean absolute error (MAE) between the PPG and reference HR exceeded 10 bpm. Activity labels were assigned if a diary-reported activity related to structured PA (e.g., rehabilitation training, walking, running, cycling) was present, or if the HR-based prominence peak detection algorithm detected a significant HR increase corresponding to activity. Several features capturing HR dynamics (including HR values, Savitsky-Golay filter outputs, Z-score values, and prominence characteristics) were calculated to provide the models with informative inputs for distinguishing between artefacts and true activity-related HR changes.

Models were trained using data from 75% of the PPG-compatible participants (randomly selected) and validated through 5-fold cross-validation (80:20 split within each fold). All remaining data, consisting of the remaining 25% of PPG-compatible participants and all PPG-incompatible participants, were retained separately as independent test data. In the present study, we focus model evaluation on the PPG-compatible testing subset, given their data quality and clinical relevance. The classifier type, a key determinant of model performance, was optimised using area under the receiver operating characteristic curves (ROC-AUC) to account for class imbalances (e.g., artefacts as a minority class). This resulted in the selection of balanced bagging with random forest for both models.[24] Additionally, parameter tuning was refined by integrating a pipeline with grid search principles, which tested multiple combinations of values for both prominence and classifier parameters and selected the configuration yielding the highest F1-score.

Training both models took approximately 5 h on an Intel(R) Core(TM) i5–9500 T CPU with 8 GB RAM. Application, including preprocessing and predictions from both models, required only 5 s to process 24 h of HR data from a single participant, making the ARP both lightweight and computationally efficient.

Combining prediction scores to label unreliable data and achieve clinically relevant HR data

Applying the two trained models to the testing data resulted in artefact and activity prediction scores (i.e., probabilities) between 0 and 100% for every datapoint, where 0% indicates no predicted presence of the class (artefact or activity), and 100% indicates full predicted presence. The goal was to combine these prediction outputs to label data as unreliable and to achieve clinically relevant HR data with a balance between artefact detection and activity preservation.

The optimal combination of prediction thresholds was determined on the PPG-compatible testing dataset for daily life and exercise conditions. Thirty-five combinations with the following format were tested: 'data are labelled as unreliable when the artefact probability $> X$% while activity probability $< Y$%', where X ranged from 0.5–0.9 and Y from 0.3–0.9. From these, we first selected all combinations achieving a minimum sensitivity to detect artefacts of at least 30% during both daily life and exercise conditions, as a pragmatic lower bound to ensure clinical utility.

For each remaining combination of prediction thresholds, datapoints labelled as unreliable were removed and interpolated with the mean of adjacent reliable values (maximum interpolation window: 10 min). We evaluated the clinical validity of each combination of prediction thresholds by calculating the Antwerp Activity Index (AAI), an HR-based PA score described in section Calculation of the Antwerp Activity Index (AAI), on the cleaned data and comparing it to the reference AAI obtained from the Polar device (as gold standard). The combination that led to the highest agreement with the reference AAI (the highest Pearson's correlation and the lowest MAE) was selected as optimal. Additionally, AAI discrepancy analysis was performed to evaluate the number and type of large deviations (>20 points) between the PPG- and reference-based AAI. Particular attention was given to clinically relevant discrepancies, defined as cases where one AAI score equalled zero while the other

exceeded 20 points, i.e., complete underestimation (PPG AAI = 0, reference AAI > 20) or overestimation (reference AAI = 0, PPG AAI > 20). These specific scenarios may indicate a complete miss or false detection of activity, which could have practical consequences for patient motivation or clinical interpretation.

Calculation of the Antwerp Activity Index (AAI). The AAI score quantifies PA using HR data and provides an interpretable metric for both users and health professionals, where higher exercise intensities yield higher AAI scores. The AAI calculation process is illustrated in Supplemental Figure 2, with the corresponding mathematical formulation provided in Supplemental Table 2. We adapted the concept of the Personal Activity Intelligence (PAI) algorithm originally developed by the Cardiac Exercise Research Group (CERG) at the Norwegian University of Science and Technology (NTNU).[25] The calculation begins with the determination of an activity score. This involves normalising HR data against an individualised threshold HR, above which values start contributing to the AAI. We defined the threshold HR as HR_{rest} + 40% of heart rate reserve (HRR), which corresponds to the definition of moderate intensity exercise according to the ESC Guidelines of Sports Cardiology 2020.[26] We calculated HR_{rest} daily as the median of the 30 lowest HR values recorded between 7 a.m. and 11 p.m., while HR_{max} is based on a maximal cardiopulmonary exercise test at onset (which may be repeated later). HR values below the threshold HR are excluded from further analysis, reflecting the principle that very light activity offers limited cardiovascular benefit.[27] A non-linear transformation is applied for HR values above the threshold, which reflects the increasing physiological load at higher intensities. The transformation employs gender-specific constants: c_1 was adopted from the original PAI algorithm,[25] while c_2 was optimised in our cohort to maximise the correlation with measured VO_2 max change as measured by cardiopulmonary exercise tests at the beginning and end of the study. The resulting transformed values are integrated over time to compute the activity score. Finally, the activity score is passed through a logarithmic scaling function, resulting in the final AAI score. This transformation includes a saturation effect: initial periods of moderate-to-vigorous PA lead to greater increases in AAI, while gains flatten at higher loads.[27,28] The scaling constants (c_3 and c_4) were determined by fitting a logarithmic function such that an AAI score of 100 corresponds approximately to a 10% improvement in VO_2 max, as observed in our population. This procedure ensures that AAI remains a meaningful, personalised score across different fitness levels. In this study, AAI values were calculated separately for both the reference HR data (from the chest strap) and for the ARP-processed PPG data. The agreement between both versions of the AAI was used to evaluate the clinical validity of the ARP.

Statistical analysis

All statistical analyses were performed using SPSS Statistics version 29 (IBM Corp) and Python version 3.9.

As described in section Training and optimisation of the ARP: artefact and activity models, both the artefact and activity models were trained on 75% of the PPG-compatible participants using 5-fold cross-validation. To assess the performance of the ARP, several confusion matrices were calculated using the independent test data, i.e., the remaining 25% of PPG-compatible participants. Confusion matrices were stratified by condition (daily life vs. exercise), where 'daily life' corresponds to 24 h measuring periods and 'exercise' includes all PA registered during the 12-week monitoring period. From these matrices, we derived common performance metrics such as accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

To further evaluate the benefit of the ARP, accuracy metrics were computed for the PPG data before and after artefact removal (in the Supplementary materials). Pearson's correlation coefficient (r) evaluated linear agreement between PPG HR and reference HR (Polar). Device error was quantified using MAE (bpm) and MAPE (%), with MAPE <10% considered clinically acceptable.[21–23] Error classification using raw percentage error was also used to determine undershooting ($\leq -10\%$) and overshooting ($\geq +10\%$).

Descriptive statistics for baseline demographic and clinical characteristics were reported as means (± standard deviation) for continuous variables and as counts (percentages) for categorical variables. Normality of continuous variables was assessed using the Shapiro–Wilk test and visual inspection of histograms. Group-level comparisons for baseline characteristics were conducted between the three main study groups (CR patients, coached individuals, and non-coached individuals), between PPG-compatible and PPG-incompatible participants, as well as between participants included in the training vs testing datasets. As normality could not be assumed in all groups, non-parametric Kruskal–Wallis tests were used for between-group comparisons of non-paired continuous variables (e.g., age and weight). For categorical variables (e.g., gender and skin type), chi-square tests were used to compare distributions across groups. If expected cell counts were below 5, Fisher's Exact test was performed. All p-values were two-sided, and a significance level of $p < 0.05$ was used throughout. Despite some non-normal distributions, descriptive statistics were reported as means and standard deviations when preferred for interpretability and consistency with clinical reporting practices.

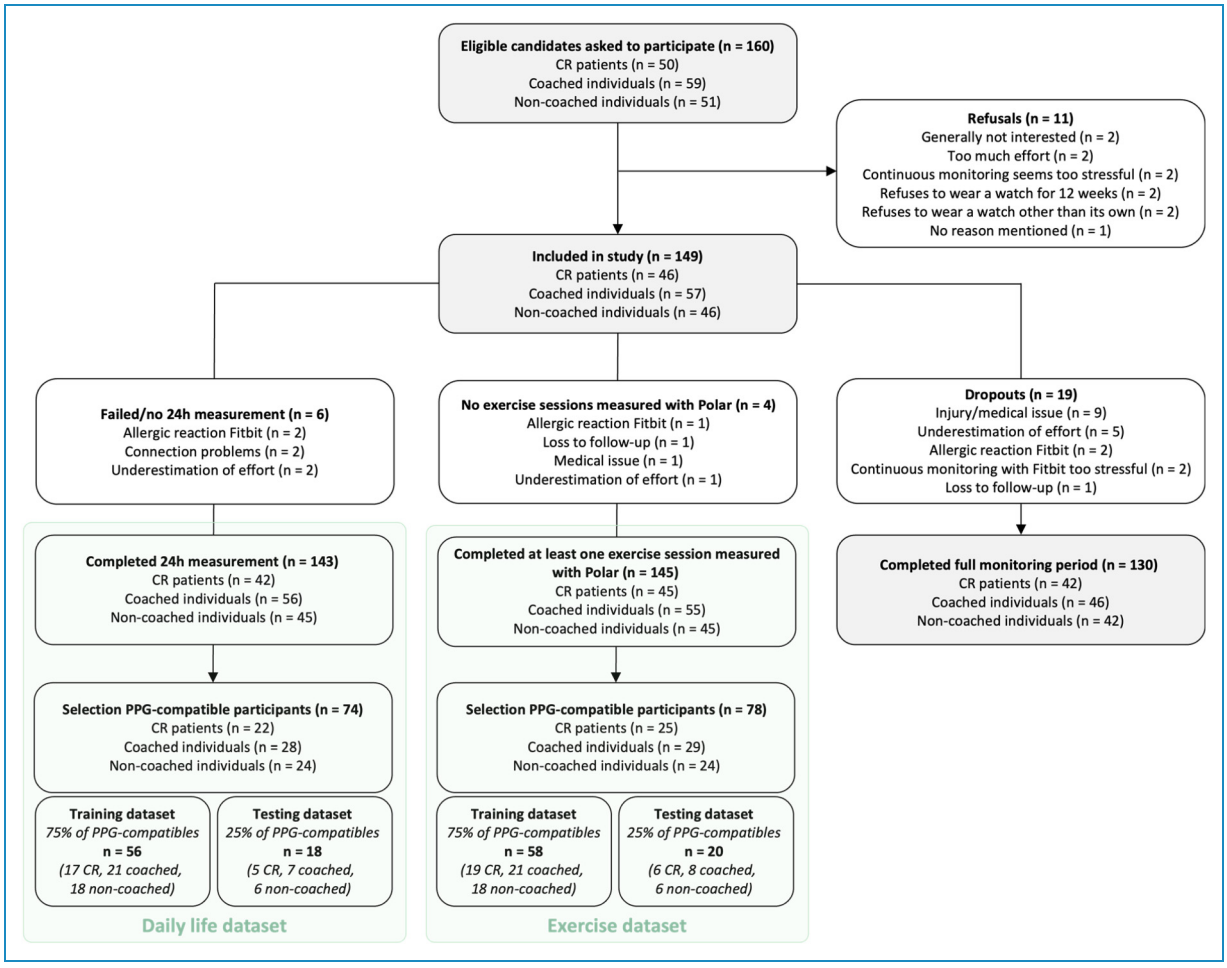

Figure 1. Study flow chart.
CR: cardiac rehabilitation. PPG compatibility is defined as ≥70% accurate training time (mean absolute percentage error <10%).

Results

Demographics & PPG-compatibility analysis

A total of 161 participants were invited to take part in the study, of which 149 were enrolled (CR patients: $n = 46$; coached individuals: $n = 57$; non-coached individuals: $n = 46$) (Figure 1). While the two healthy groups (coached and non-coached) were demographically similar, CR patients were significantly older, had higher BMI, more cardiovascular comorbidities (e.g., hypertension, hypercholesterolemia), and more frequent medication use, most notably beta blockers (Supplemental Table 3). Over a mean follow-up of 12.6 ± 3.5 weeks, participants recorded an average of 31.3 ± 19.8 exercise sessions using the Polar chest strap, with comparable numbers across CR patients (32.7 ± 16.1), coached (31.3 ± 22.0), and non-coached individuals (29.9 ± 20.5) ($p = .79$).

PPG-compatibility analysis was feasible in 145 of the 149 participants, as four lacked sufficient training data. For most, the analysis was based on three training sessions as intended ($n = 141$), and in a few cases on one ($n = 3$) or two ($n = 1$) sessions. Seventy-eight participants (53.8%) were classified as PPG-compatible and 67 (46.2%) as PPG-incompatible. Table 1 summarises the baseline characteristics of both groups. No significant differences were observed in age, gender, skin type, or cardiovascular risk factors, although BMI was significantly higher in the PPG-compatible group (26.4 ± 3.8 vs. 24.9 ± 3.6 kg/m^2, $p = .02$).

Of the 78 PPG-compatible participants, 58 (75%) were randomly allocated to the model training dataset and 20 (25%) to the independent testing dataset. Baseline characteristics were balanced between training and testing datasets (Supplemental Table 4). Unless otherwise specified, results presented in the following sections are based on the PPG-compatible testing group.

Performance of the artefact and activity models

The training dataset consisted of a total of 963,006 daily life and 3,181,648 exercise datapoints. This difference reflects

Table 1. Baseline characteristics of the study population, split for PPG-compatibility.

	All participants ($n = 149$)	PPG-compatible participants ($n = 78$)	PPG-incompatible participants ($n = 67$)	p-value (PPG-compatible vs PPG-incompatible)
Group, *n* (%)				0.96
CR patients	46 (30.9%)	25 (32.1%)	20 (29.9%)	
Coached individuals	57 (38.3%)	29 (37.2%)	26 (38.8%)	
Non-coached individuals	46 (30.9%)	24 (30.1%)	21 (31.3%)	
Age (years)				0.32
Mean ± SD	48.7 ± 13.5	49.8 ± 13.3	47.4 ± 13.8	
Range	20–78	20–78	24–75	
Gender, *n* (%)				0.69
Male	109 (73.2%)	56 (71.8%)	51 (76.1%)	
Female	40 (26.8%)	22 (28.2%)	16 (23.9%)	
Weight (kg)				0.19
Mean ± SD	81.1 ± 14.2	82.6 ± 13.8	79.4 ± 14.0	
Range	55.0–138.4	57.1–138.4	55.0–129.8	
BMI (kg/m^2)				0.02
Mean ± SD	25.7 ± 3.8	26.4 ± 3.8	24.9 ± 3.6	
Range	17.4–39.9	18.7–39.9	17.4–35.6	
Skin type[a], *n* (%)				0.88
Type 1	16 (11.0%)	9 (11.7%)	7 (10.8%)	
Type 2	81 (55.5%)	42 (54.5%)	35 (53.8%)	
Type 3	44 (30.1%)	24 (31.2%)	20 (30.8%)	
Type 4	4 (27.4%)	2 (2.6%)	2 (3.1%)	
Type 5 & 6	1 (0.7%)	0 (0.0%)	1 (1.5%)	
Medical history, *n* (%)				
Atrial fibrillation	12 (8.1%)	8 (10.3%)	4 (6.0%)	0.53
HFrEF	1 (0.7%)	1 (1.3%)	0 (0.0%)	1.00
Coronary artery disease	34 (22.8%)	21 (26.9%)	12 (17.9%)	0.28
Without PCI/CABG	4 (2.7%)	2 (2.6%)	2 (3.0%)	1.00

(continued)

Table 1. Continued.

	All participants ($n = 149$)	PPG-compatible participants ($n = 78$)	PPG-incompatible participants ($n = 67$)	p-value (PPG-compatible vs PPG-incompatible)
With PCI	26 (17.4%)	16 (20.5%)	9 (13.4%)	0.37
With CABG	8 (5.4%)	5 (6.4%)	3 (4.5%)	0.89
TIA	1 (0.7%)	1 (1.3%)	0 (0.0%)	1.00
CVA	4 (2.7%)	3 (3.8%)	1 (1.5%)	0.72
Myocardial infarction	7 (4.7%)	4 (5.1%)	3 (4.5%)	1.00
Valve disease	12 (8.1%)	8 (10.3%)	4 (6.0%)	0.51
Vascular disease	11 (7.4%)	6 (7.7%)	5 (7.5%)	1.00
Pulmonary embolism	1 (0.7%)	0 (0.0%)	1 (1.5%)	0.94
Peripheral embolism	1 (0.7%)	1 (1.3%)	0 (0.0%)	1.00
Cardiovascular risk factors, *n (%)*				
Hypercholesteremia	49 (32.9%)	31 (39.7%)	18 (26.9%)	0.15
Hypertension	29 (19.5%)	17 (21.8%)	11 (16.4%)	0.54
Diabetes (type 2)	1 (0.7%)	1 (1.3%)	0 (0.0%)	1.00
Smoking status				0.73
Previous	39 (26.4%)	20 (25.6%)	18 (27.3%)	
Current	6 (4.1%)	3 (3.8%)	3 (4.5%)	
Never	103 (69.6%)	55 (70.5%)	45 (68.2%)	
Medication use, *n (%)*				
Rate control	29 (19.5%)	20 (25.6%)	9 (13.4%)	0.10
Beta blocker	20 (13.4%)	14 (17.9%)	6 (9.0%)	0.19
Calcium antagonist	13 (8.7%)	9 (11.5%)	4 (6.0%)	0.38
Rhythm control	5 (3.4%)	4 (5.1%)	1 (1.5%)	0.46
Anti-coagulantia	16 (10.7%)	10 (12.8%)	6 (9.0%)	0.64
Anti-platelet	40 (26.8%)	25 (32.1%)	14 (20.9%)	0.19
ACE-inhibitor	18 (12.1%)	12 (15.4%)	6 (9.0%)	0.36
Angiotensin II receptor blocker	6 (4.0%)	4 (5.1%)	2 (3.0%)	0.82
Centrally acting antihypertensives	0 (0.0%)	0 (0.0%)	0 (0.0%)	NA

(continued)

Table 1. Continued.

	All participants ($n = 149$)	PPG-compatible participants ($n = 78$)	PPG-incompatible participants ($n = 67$)	p-value (PPG-compatible vs PPG-incompatible)
Statin	47 (31.8%)	30 (38.5%)	17 (25.4%)	0.15
Hypolipaemic non-statin	27 (18.2%)	18 (23.1%)	8 (11.9%)	0.14
Gastric acid secretion inhibitor	21 (14.1%)	13 (16.7%)	8 (11.9%)	0.57
Diuretic	6 (4.0%)	4 (5.1%)	2 (3.0%)	0.82

CR: cardiac rehabilitation; BMI: body mass index; PCI: percutaneous coronary intervention; CABG: coronary artery bypass grafting; CVA: cerebrovascular accident; TIA: transient ischemic attack; HFrEF: heart failure with a reduced ejection fraction (left ventricular ejection fraction, LVEF, <40%).
[a]Skin type was determined according to the Fitzpatrick classification, ranging from skin type 1 (pale white skin) to type 6 (dark brown or black skin).
PPG-compatibility analysis was performed in 145 out of 149 participants. Four participants were excluded due to the absence of training data.
PPG-compatible participants achieved ≥70% accurate training time (mean absolute percentage error <10%), while PPG-incompatible participants scored below this threshold.
Missing data were present for skin type ($n = 3$), valve disease ($n = 1$), smoking status ($n = 1$), centrally acting antihypertensives ($n = 1$), statin use ($n = 1$), and non-statin lipid-lowering drugs ($n = 1$).

the study protocol: reference chest strap data (needed for model training) were collected during a single 24-h daily life period and during all exercise sessions throughout the 12-week monitoring period. The PPG compatible testing dataset included 311,904 datapoints collected during daily life (mean per participant: 17,328), of which 85,976 were labelled as artefacts (i.e., reference–PPG error >10 bpm) and 40,412 as activity (based on diaries and HR peak detection). During exercise, 680,276 datapoints were collected, including 354,084 artefacts and 384,068 activities (mean per participant: 34,014). Artefacts were more prevalent during exercise (52.1%) than during daily life (27.6%).

The automatic artefact model (Figure 2(A) and (B)) detected over half of the artefacts in daily life conditions (58.5% sensitivity) while rarely misclassifying correct data (91.3% specificity), resulting in an overall accuracy of 82.2%. The PPV and NPV were 71.9% and 85.2%, respectively. During exercise, the artefact model identified over three-quarters of artefacts (76.6% sensitivity) while still limiting incorrect removals (74.9% specificity, 75.8% accuracy). The PPV and NPV were 76.8% and 74.7%, respectively.

The activity model (Figure 2(C) and (D)) reached a sensitivity of 74.4% during daily life, correctly identifying most active periods. Specificity was 84.4%, while PPV remained relatively low (41.5%) due to frequent false positives. Accuracy was 83.1%, and NPV was high at 95.7%. During exercise, the model correctly detected 88.4% of activity episodes (sensitivity), but misclassified many inactive datapoints as activity (40.1% specificity). The PPV and NPV were 65.7% and 72.6%, respectively, and the overall accuracy reached 67.3%.

Minor differences were observed across participant types (exploratory overview in Supplemental Tables 6 and 7): CR patients generally showed slightly higher artefact model accuracy (+7.7%) and lower activity model accuracy (−5.3%) than coached individuals, with non-coached individuals performing intermediately.

Determination of the optimal combination rule

To identify the most effective rule for labelling unreliable data and activity data, we tested 35 combinations of artefact and activity probability thresholds. Five threshold combinations met the predefined sensitivity criterion of ≥30% for artefact detection in both daily life and exercise conditions (Supplemental Table 5). Among these, the combination labelling data as unreliable when artefact probability >50% and activity probability <70% consistently showed the best agreement between PPG-derived AAI and reference AAI.

This agreement is visualised in scatter and Bland–Altman plots in Figure 3, which allow visual identification of both large and clinically relevant discrepancies. For this optimal threshold rule, large AAI discrepancies, defined as differences of >20 points and shown as crosses in the plots, were limited to 5.7% of daily life and 8.3% of exercise measurements (Figure 4(A)). Clinically relevant discrepancies, where one AAI score was 0 and the other exceeded 20, occurred in only 5.7% (daily life) and 1.4% (exercise), with the lowest underestimation rates and minimal overestimation (maximum 0.4%) (Figure 4(B)). These discrepancies are identifiable on the zero lines in the scatter plots of Figure 3: horizontally for underestimation (PPG AAI = 0), and vertically for overestimation (reference AAI = 0).

The Bland–Altman plots further illustrate whether these discrepancies fall within the limits of agreement (LOA). Under the optimal rule, only eight clinically relevant

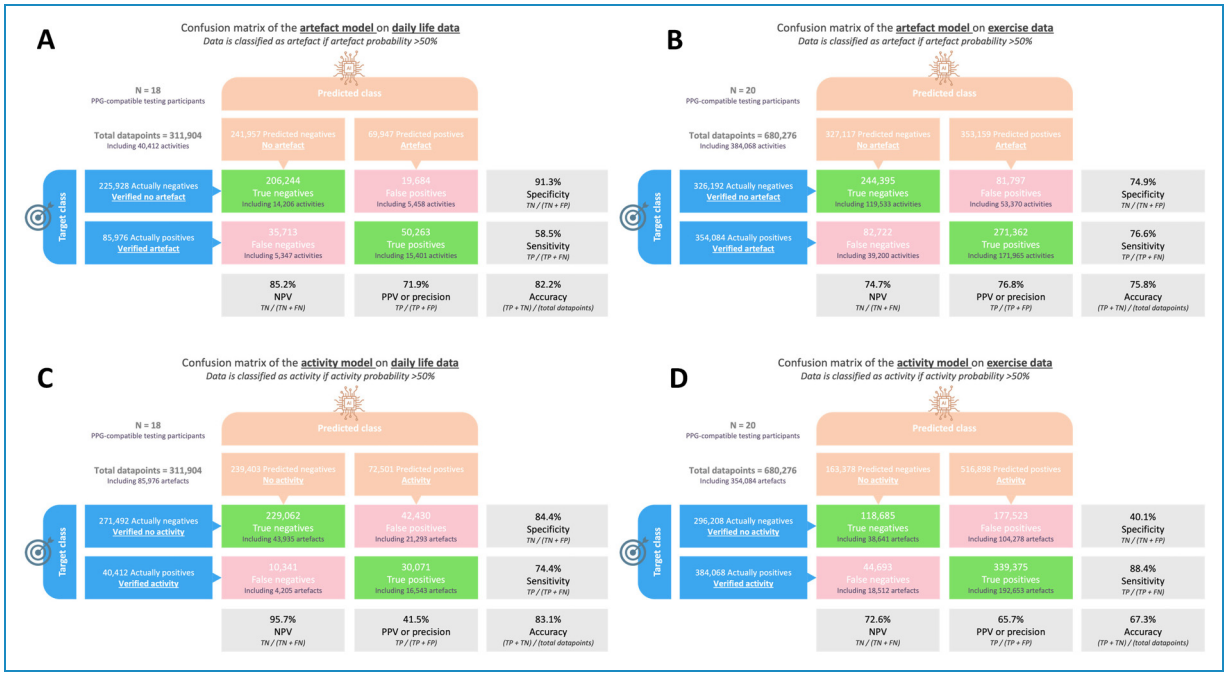

Figure 2. Performance of the artefact and activity model on the PPG-compatible testing dataset, in daily life (A, C) and exercise (B, D) conditions.
PPV: positive predictive value; NPV: negative predictive value; TP: true positives; FP: false positives; TN: true negatives; FN: false negatives.

discrepancies (five underestimations and three overestimations) out of 45 total AAI discrepancies exceeded the LOA, and are indicated with thin blue lines in Figure 3(A). These eight originated from seven different participants, each of whom had the majority of their datapoints within the LOA (ranging from 16 to 54 measurements within the LOA per participant).

Finally, this optimal combination rule also resulted in the highest Pearson's correlation coefficients of $r = .96$ (daily life) and $r = .89$ (exercise), and the lowest MAEs of 5.2 and 6.7 AAI points, respectively (Figure 5(A) and (B)).

Final performance after applying the optimal combination of prediction thresholds

After applying the optimal threshold combination rule, the ARP still had a sensitivity to detect 39.1% of all artefacts during daily life and 31.9% during exercise (Supplemental Figure 3). The data incorrectly classified as unreliable (i.e., false positives) contained only a small proportion of true activity: 4.4% (1,774 out of 40,412) during daily life and 5.3% (20,524 out of 384,068) during exercise. This illustrates that the ARP effectively preserves activity-relevant information while filtering out artefacts. This principle is also illustrated in the ARP decision-making examples (Figure 6). In one case, a clear PPG overshooting was correctly rejected as unreliable. In another, undershooting HR values during activity (e.g., cycling) were preserved due to high activity probability scores, ensuring that activity information was not lost.

We also compared HR accuracy metrics before and after applying the ARP (Supplemental Figure 4). The ARP consistently improved agreement with the reference across all conditions. Supplemental Tables 6 and 7 indicate that final ARP performance varied slightly across testing subgroups. For example, during daily life, accuracy ranged from 69.3% in coached individuals to 81.2% in CR patients.

Discussion

This study is the first to validate a novel ML approach for detecting unreliable PPG-derived HR data from consumer-grade wrist-worn devices. By combining independently trained models, for artefact detection and activity recognition, the ARP excludes unreliable data with minimal loss of activity-relevant data, hence correctly identifying activity-related HR fluctuations. ARP performance was evaluated in relation to the AAI, an HR-based PA score, to assess the clinical information derived from the HR data.

Optimised artefact detection in real-world settings for clinically meaningful activity quantification

Whereas previous artefact detection methods rely on raw PPG waveforms, tied to research-grade setups or

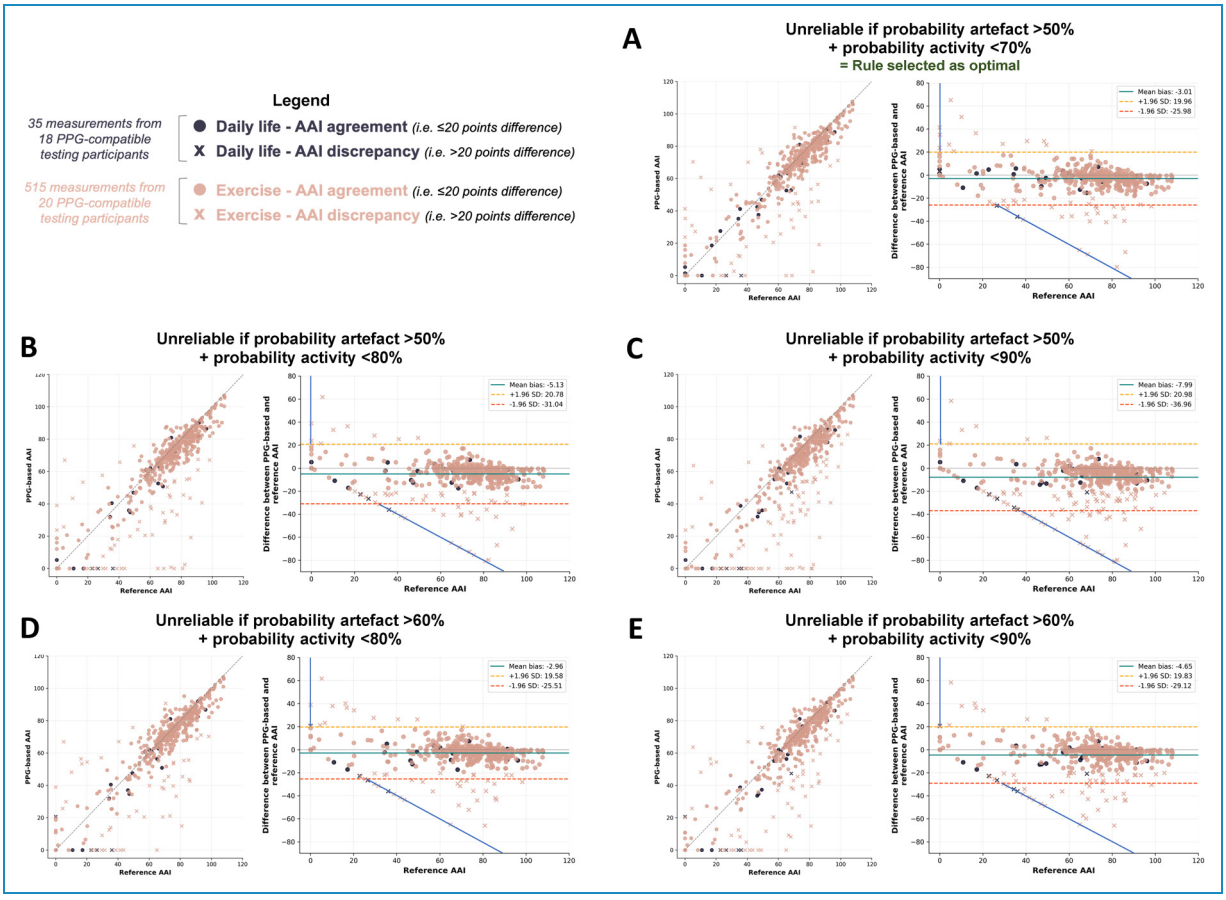

Figure 3. Scatter and Bland–Altman plots for PPG-based and reference AAI across five threshold combinations, evaluated in the PPG-compatible testing dataset under both daily life and exercise conditions.
Scatter plots show the relation between reference AAI (*x*-axis) and PPG-based AAI (*y*-axis), with visual distinction between daily life and exercise, and between datapoints with ≤20-point difference ("AAI agreement") and those with >20-point difference ("AAI discrepancies"). Bland–Altman plots depict the differences (PPG-based AAI – reference AAI) against the reference AAI, with mean bias and limits of agreement (±1.96 SD) indicated. Blue lines in the Bland–Altman plots indicate the location of clinically relevant discrepancies (i.e., difference >20 AAI points and of the two AAI values is zero) that fall outside the limits of agreement. Threshold combination rules were selected based on a minimum artefact detection sensitivity of 30% during both daily life and exercise conditions (see Supplementary Table 4).

proprietary commercial software,[10–13] our approach operates on derived HR values from a consumer-grade device. This makes the ARP suited for real-world applications in clinical and outpatient contexts, where raw PPG signals are typically unavailable and only device generated HR data can be accessed. As noted in our prior pilot study,[16] this approach bridges a key gap between academic innovation and practical deployment.

The current version of the ARP underwent several technical refinements over our initial report[16]: an expanded training dataset and independent testing dataset, improved target labelling, optimised model selection (balanced bagging with random forest), and the use of grid search for hyperparameter tuning. In addition, a core innovation of this study lies in the clinical context of our artefact labelling. Rather than relying solely on artefact detection or technical classification metrics, we optimised the ARP using a combined artefact and activity model, guided by its impact on PA quantification via the AAI score.

Using only the artefact model resulted in high sensitivity (58.5% during daily life, 76.6% during exercise), but also misclassified a notable proportion of activity-relevant data as unreliable (13.5% and 13.9%, respectively). Integrating the activity model and optimising thresholds based on AAI agreement reduced this loss to 4.4% and 5.3%, respectively, albeit at the cost of lower final sensitivity (39.1% and 31.9%). This trade-off is crucial in clinical settings, where preserving PA-relevant signals is often more important than maximising artefact removal.

Although differences between the combination rules were small, the optimal one showed the strongest correlation between PPG and reference AAI scores ($r = .96$ for daily life and $r = .89$ for physical activity) and the lowest MAEs (5.2 and 6.7 points, respectively). We identified clinically

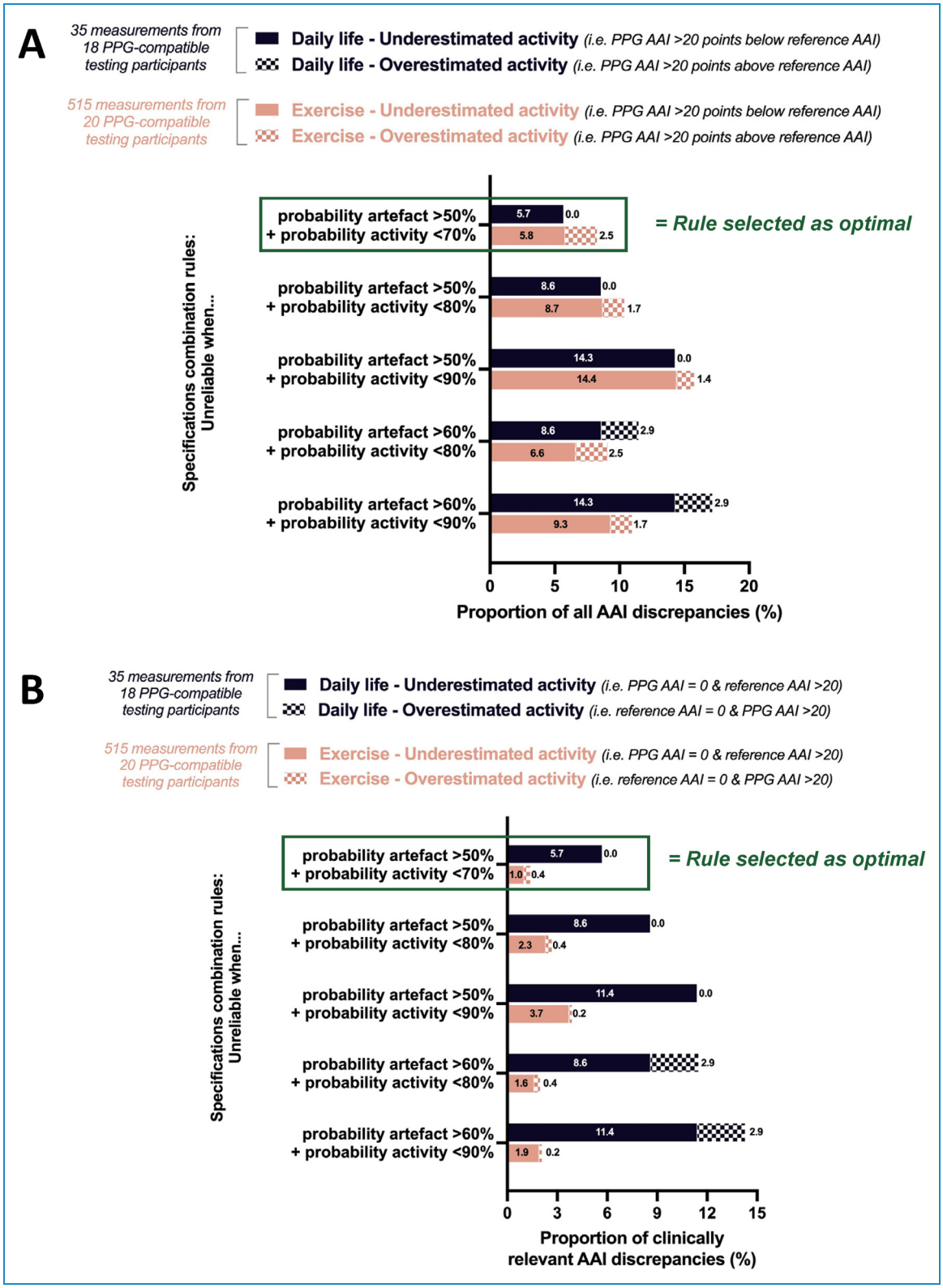

Figure 4. Proportion of AAI discrepancies between PPG-based and reference AAI across five threshold combinations, evaluated in the PPG-compatible testing dataset under both daily life and exercise conditions. (A) All discrepancies >20 AAI points and (B) Clinically relevant discrepancies where one of the AAI scores was zero.

Threshold combination rules were selected based on a minimum artefact detection sensitivity of 30% during both daily life and exercise conditions (see Supplemental Table 4). For each rule, data classified as unreliable were removed, interpolated, and subsequently used to calculate a PPG-based AAI. AAI-discrepancies were categorised as underestimated or overestimated activity, according to the legend.

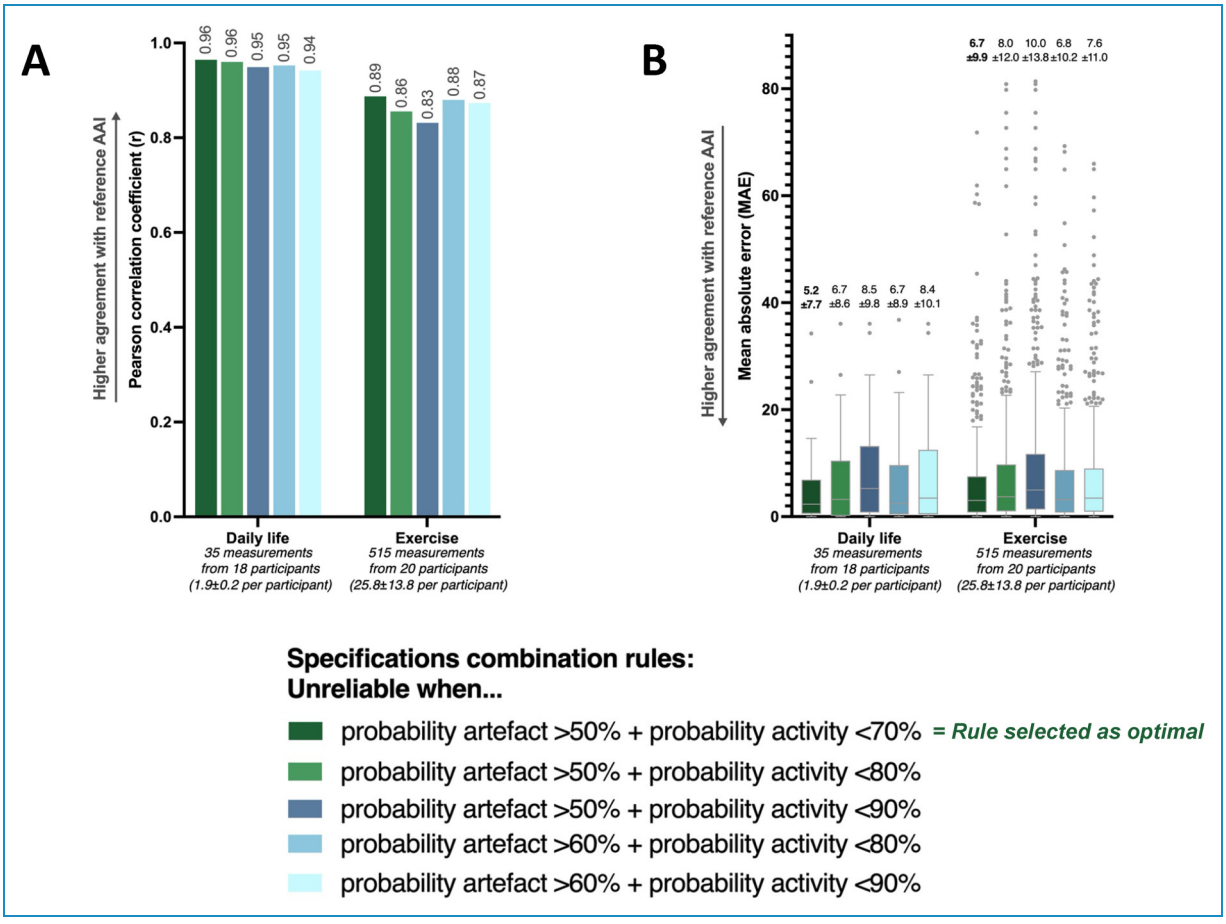

Figure 5. Agreement between PPG-derived and reference AAI across five rules for combining artefact and activity detection thresholds, evaluated in the PPG-compatible testing dataset under both daily life and exercise conditions. (A) Pearson's correlation coefficients (*r*) and (B) Mean absolute errors (MAEs, in AAI points).
Threshold combination rules were selected based on a minimum artefact detection sensitivity of 30% during both daily life and exercise conditions (see Supplemental Table 4). For each rule, data classified as unreliable were removed, interpolated, and subsequently used to calculate a PPG-based AAI. Agreement with the reference AAI was assessed using Pearson's correlation (panel A) and mean absolute error (MAE, panel B). MAE (panel B) is represented via boxplots due to non-normality, with interquartile range (IQR), median, whiskers extending to 1.5× IQR, and dots representing outliers. For interpretability and consistency with clinical literature, the mean ± standard deviation (SD) is reported above each box. Full threshold specifications per rule are shown in the colour legend below.

relevant discrepancies in only 8 of 45 AAI discrepancies (18%), all isolated within otherwise accurate profiles. These involved full under- or overestimation, which could demotivate or falsely reassure patients,[29] but were rare and not systematic, supporting ARP's clinical reliability. We have shown that the ARP approach, which can be applied to other devices, populations, and larger cohorts, is not only feasible but also meaningful for clinical applications. Applicability will need to be tested in clinical trials.

Our results are not directly comparable to previous work, given the substantial methodological differences. Previous studies that report high sensitivities (85.5%-95.7%) and accuracies (84.0%-95.3%) typically rely on raw PPG waveform data, controlled data acquisition settings, or additional sensor inputs such as inertial data (e.g., accelerometry or gyroscopy).[10,12,13] Still, our artefact model achieved 76.6% sensitivity and 75.8% accuracy during exercise, approaching the performance of conventional methods.[10,12,13] For context, a 24-h Holter ECG showed only 0.096% artefacts after expert review, of which standard Holter software detected just 29%, highlighting our model's strong performance under noisier PPG conditions.

The role of PPG-compatibility and ARP performance across user groups

While the ARP improves HR data reliability, its performance depends on the quality of the input signal, with PPG compatibility as a key factor. Consistent with our earlier work,[15] 53.8% of participants were classified as PPG-compatible. Interestingly, these individuals had higher

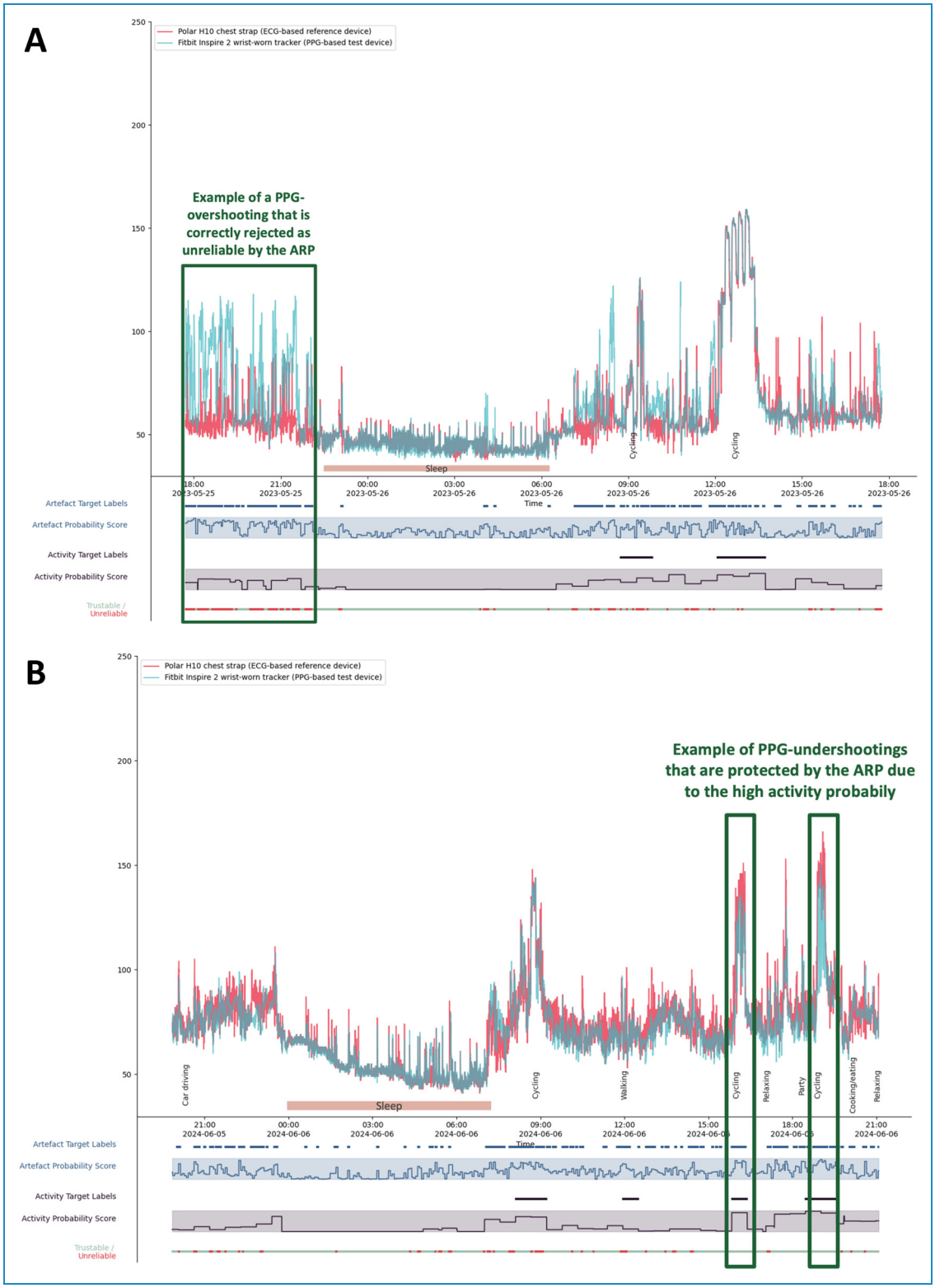

Figure 6. Visual examples illustrating the ARP decision process using the optimal combination rule (artefact probability >50% and activity probability <70%) in daily life recordings.

ARP: artefact removal procedure; PPG: photoplethysmography; ECG: electrocardiogram. Graphs show the original heart rate recordings (PPG vs. reference), target labels for artefact and activity detection, model probabilities (visualised within bars ranging from 0% at the bottom to 100% at the top), and the resulting ARP decision (green = trustable, red = unreliable), based on the optimal combination rule: data are labelled as unreliable when artefact probability >50% and activity probability <70%.

BMI, contradicting common assumptions that higher BMI reduces PPG signal quality.[30] Similar paradoxical findings in other studies suggest a more complex relationship between BMI and PPG quality.[16,31] No other demographic or clinical factors consistently predicted compatibility.

This lack of a clear predictive profile highlights the importance of conducting a PPG-compatibility check prior to clinical use. Clinicians and researchers should avoid assuming the reliability of wrist-worn PPG data without verification.

The ARP performed comparably across subgroups, with CR patients showing slightly better artefact detection but marginally lower activity classification accuracy. In PPG-incompatible individuals, accuracy improved somewhat with ARP, but MAPE remained above the 10% clinically acceptable thresholds (not shown), confirming that even ARP cannot compensate for unreliable data in PPG-incompatible individuals.[16] Thereby reaffirming the importance of preselecting for PPG-compatibility.

Clinical integration and future applications

The ARP, when paired with daily AAI extraction, offers a pathway for continuous, accurate, and objective HR-based PA monitoring using commercial devices. As the ARP is lightweight, explainable, and computationally efficient, it is well-suited for integration into mobile health platforms.[32,33] Therefore, it could facilitate more personalised patient feedback, support remote follow-up, and guide training in CR settings, including ambulatory CR after a hospital-based phase, where structured hospital supervision is no longer available[34,35] and adherence to PA recommendations often declines.[36,37]

While the present study focused on optimising and validating the ARP, we are currently designing a multicentre randomised controlled clinical trial to evaluate the combined ARP–AAI system after a hospital-based CR phase. Patients will wear a wrist-based PPG device that passively collects and securely transmits HR data. The ARP processes this data automatically, after which the cleaned output is used to compute the AAI. This score is visualised for both patients (via a mobile app) and clinicians (via a dashboard), supporting personalised feedback. The trial will assess the fully automated system's ability to improve clinical outcomes, as well as adherence to WHO PA recommendations.[38] As more data will be collected in this trial, the selected ARP combination thresholds may be re-evaluated and refined to maintain optimal performance.

Limitations

Several limitations should be acknowledged. First, we used data from one commercial PPG device (i.e., Fitbit Inspire 2) whose internal signal processing is not transparent. While our method applies to any wearable providing HR data export and offering similar HR sampling resolution, generalisability remains unconfirmed. Second, while our training dataset was large in terms of datapoints (i.e., 4,144,654) and we applied no exclusions based on sex or skin tone, it could be expanded to better capture subgroups such as women and individuals with darker skin.[30,39] Third, the AAI, based on the retrospectively validated PAI framework, has not yet been prospectively validated, though a study to address this is underway (as discussed at the end of section Clinical integration and future applications). Next, as rhythm monitoring was not done in this study, undetected AF episodes cannot be excluded, and the impact of rhythm irregularity on ARP performance could not be assessed here. In addition, as only 53.8% of participants were classified as PPG-compatible, clinical implementation requires a brief pre-use screening step to identify suitable users. Finally, although we standardised device application (e.g., advice on strap position and tightness), we did not measure all contextual factors that may affect signal quality, such as strap tightness, vascular perfusion, or temperature.[30,39,40] Future studies should consider capturing these variables and evaluating their influence on real-world performance.

Conclusions

This study validated a lightweight and explainable ML based ARP for consumer grade PPG devices that operates directly on device generated HR data, rather than on raw PPG waveforms. The proposed method allows accurate and continuous HR-based activity monitoring by effectively removing artefacts while preserving clinically relevant PA data, and is well-suited for integration into mobile health platforms. Future trials will prospectively assess its ability to improve PA feedback, support adherence, and contribute to improved clinical outcomes.

Acknowledgements

The authors would like to sincerely thank all study participants for their time and effort in contributing to this research. We also extend our gratitude to the nurses, physicians, physiotherapists, and other colleagues who assisted in recruiting participants and facilitating data collection.

ORCID iDs

Paulien Vermunicht https://orcid.org/0000-0001-5922-2095
Christophe Buyck https://orcid.org/0009-0005-4214-9932
Sebastiaan Naessens https://orcid.org/0009-0006-9395-2271
Wendy Hens https://orcid.org/0000-0002-9881-6248
Katsiaryna Makayed https://orcid.org/0009-0004-2704-9530
Lien Desteghe https://orcid.org/0000-0001-8641-4658

Ethical considerations

The study was approved by the Ethics Committee of University Hospital Antwerp (UZA) and the University of Antwerp, and

conducted in accordance with the Declaration of Helsinki (EC reference: 2023–5241, BUN: B3002023000046).

Consent to participate

All participants in this study provided written informed consent prior to their inclusion.

Consent for publication

Not applicable. This study reports group-level data, and no individual or identifiable data are included in this manuscript.

Author contributions

Conceptualisation: all authors; Formal analysis: Paulien Vermunicht; Investigation (i.e., conducting study): Paulien Vermunicht, Christophe Buyck, Sebastiaan Naessens, Wendy Hens, Johan Roeykens, Koen De Deckere; Software: Juan Sebastian Piedrahita Giraldo, Katsiaryna Makayed, Saartje Herman, Kris Laukens, Paulien Vermunicht; Writing–original draft: Paulien Vermunicht; Writing–review and editing: all authors.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: this study was supported by the FWO (Fonds Wetenschappelijk Onderzoek) senior research project 'G084023N'.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability

Raw data supporting the conclusions of this article will be made available by the authors upon request. Figures are uploaded as separate image files via the submission platform. Legends and titles for figures are grouped in this separate section of the manuscript.

Supplemental material

Supplemental material for this article is available online.

References

1. Cleven L, Krell-Roesch J, Nigg CR, et al. The association between physical activity with incident obesity, coronary heart disease, diabetes and hypertension in adults: a systematic review of longitudinal studies published after 2012. *BMC Public Health* 2020; 20. DOI: 10.1186/s12889-020-08715-4
2. Dempsey PC, Rowlands AV, Strain T, et al. Physical activity volume, intensity, and incident cardiovascular disease. *Eur Heart J* 2022; 43: 4789–4800.
3. Nes BM, Gutvik CR, Lavie CJ, et al. Personalized activity intelligence (PAI) for prevention of cardiovascular disease and promotion of physical activity. *Am J Med* 2017; 130: 328–336. 20161029.
4. Etiwy M, Akhrass Z, Gillinov L, et al. Accuracy of wearable heart rate monitors in cardiac rehabilitation. *Cardiovasc Diagn The* 2019; 9: 262–271.
5. Pasadyn SR, Soudan M, Gillinov M, et al. Accuracy of commercially available heart rate monitors in athletes: a prospective study. *Cardiovasc Diagn The* 2019; 9: 379–385.
6. Gillinov S, Etiwy M, Wang R, et al. Variable accuracy of wearable heart rate monitors during aerobic exercise. *Med Sci Sports Exerc* 2017; 49: 1697–1703.
7. Scardulla F, Cosoli G, Spinsante S, et al. Photoplethysmograhic sensors, potential and limitations: is it time for regulation? A comprehensive review. *Measurement (Mahwah N J)* 2023; 218. DOI: 10.1016/j.measurement.2023.113150
8. 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: 2600–2607.
9. Al-Zaiti SS, Alghwiri AA, Hu X, et al. A clinician's guide to understanding and critically appraising machine learning studies: a checklist for ruling out bias using standard tools in machine learning (ROBUST-ML). *Eur Heart J Digit Health* 2022; 3: 125–140. 20220412.
10. Goh CH, Tan LK, Lovell NH, et al. Robust PPG motion artifact detection using a 1-D convolution neural network. *Comput Meth Prog Biomed* 2020; 196: 105596. 20200611.
11. Vicente-Samper JM, Tamantini C, Avila-Navarro E, et al. An ML-Based Approach to Reconstruct Heart Rate from PPG in Presence of Motion Artifacts. *Biosensors (Basel)* 2023; 13 20230707. DOI: 10.3390/bios13070718.
12. Athaya T and Choi S. Evaluation of different machine learning models for photoplethysmogram signal artifact detection. *I C Inf Comm Tech Co* 2020: 1206–1208.
13. Vandecasteele K, Lázaro J, Cleeren E, et al. Artifact detection of wrist photoplethysmograph signals. In *11th International Conference on Bio-inspired Systems and Signal Processing*. Madeira, Portugal: SCITEPRESS, 2018, pp.182–189.
14. Vermunicht P, Makayed K, Meysman P, et al. Validation of polar H10 chest strap and Fitbit Inspire 2 tracker for measuring continuous heart rate in cardiac patients: impact of artefact removal algorithm. *Europace* 2023; 25. DOI: 10.1093/europace/euad122.550
15. Vermunicht P, Buyck C, Naessens S, et al. Optimisation and pre-use suitability selection for wrist photoplethysmography based heart rate monitoring in cardiac patients. *Eur Heart J – Dig Health* 2025; 6: ztaf084.DOI: 10.1093/ehjdh/ztaf084
16. Vermunicht P, Makayed K, Buyck C, et al. Continuous heart rate measurements in patients with cardiac disease: device comparison and development of a novel artefact removal procedure. *DIGITAL HEALTH* 2025; 11. DOI: 10.1177/20552076251337598
17. Technology PRa. *Polar H10 Heart Rate Sensor System*. November 11, 2019 2019.
18. Schaffarczyk M, Rogers B, Reer R, et al. Validity of the Polar H10 Sensor for Heart Rate Variability Analysis during

Resting State and Incremental Exercise in Recreational Men and Women. *Sensors (Basel)* 2022; 22 20220830. DOI: 10.3390/s22176536.
19. Merrigan JJ, Stovall JH, Stone JD, et al. Validation of garmin and polar devices for continuous heart rate monitoring during common training movements in tactical populations. *Meas Phys Educ Exerc* 2023; 27: 234–247.
20. Alzraiee AH and Niswonger RG. A probabilistic approach to training machine learning models using noisy data. *Environ Modell Softw* 2024; 179. DOI: ARTN 106133 . 1016/j.envsoft.2024.106133
21. Nelson BW and 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. 2019/03/12.
22. Chow HW and Yang CC. 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. 20200428.
23. Cardiac Monitors, Heart Rate Meters, and Alarms.
24. Vermunicht P, Buyck C, Makayed K, et al. Optimisation of artefact detection in photoplethysmography heart rate data: influence of different classifiers in machine learning models. *Eur Heart J* 2024; 45. DOI: ARTN ehae6663533 .1093/eurheartj/ehae666.3533
25. Nes BM, Gutvik CR, Lavie CJ, et al. Personalized activity intelligence (PAI) for prevention of cardiovascular disease and promotion of physical Activity. *Am J Med* 2016; 130: 328–336. 20161029.
26. Pelliccia A, Sharma S, Gati S, et al. 2020 ESC guidelines on sports cardiology and exercise in patients with cardiovascular disease. *Rev Esp Cardiol (Engl Ed)* 2021; 74: 45.
27. Franklin BA, Eijsvogels TMH, Pandey A, et al. Physical activity, cardiorespiratory fitness, and cardiovascular health: a clinical practice statement of the American society for preventive cardiology part II: physical activity, cardiorespiratory fitness, minimum and goal intensities for exercise training, prescriptive methods, and special patient populations. *Am J Prev Cardiol* 2022; 12: 100425. 20221013.
28. Garcia L, Pearce M, Abbas A, et al. Non-occupational physical activity and risk of cardiovascular disease, cancer and mortality outcomes: a dose-response meta-analysis of large prospective studies. *Br J Sports Med* 2023; 57: 979–989. 20230228.
29. Dennison L, Morrison L, Conway G, et al. Opportunities and challenges for smartphone applications in supporting health behavior change: qualitative study. *J Med Internet Res* 2013; 15: e86. 20130418.
30. Fine J, Branan KL, Rodriguez AJ, et al. Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. *Biosensors (Basel)* 2021; 11. 20210416. DOI: 10.3390/bios11040126.
31. Blok S, Piek MA, Tulevski II, et al. The accuracy of heartbeat detection using photoplethysmography technology in cardiac patients. *J Electrocardiol* 2021; 67: 148–157. 20210702.
32. Muhammad D, Ahmed I, Ahmad MO, et al. Randomized explainable machine learning models for efficient medical diagnosis. *IEEE J Biomed Health Inform*2024;PP 2024111328: 1028–1039.DOI: 10.1109/JBHI.2024.3491593
33. Ranjbarzadeh R, Dorosti S, Jafarzadeh Ghoushchi S, et al. Breast tumor localization and segmentation using machine learning techniques: overview of datasets, findings, and methods. *Comput Biol Med* 2023; 152: 106443. 20221219.
34. Giuliano C, Parmenter BJ, Baker MK, et al. Cardiac rehabilitation for patients with coronary artery disease: a practical guide to enhance patient outcomes through continuity of care. *Clin Med Insights Cardiol* 2017; 11: 1179546817710028. 20170612.
35. Bjarnason-Wehrens B, McGee H, Zwisler AD, et al. Cardiac rehabilitation in Europe: results from the European cardiac rehabilitation inventory survey. *Eur J Cardiovasc Prev Rehabil* 2010; 17: 410–418.
36. Hansen D, Dendale P, Raskin A, et al. Long-term effect of rehabilitation in coronary artery disease patients: randomized clinical trial of the impact of exercise volume. *Clin Rehabil* 2010; 24: 319–327. 20100222.
37. Moore SM, Charvat JM, Gordon NH, et al. Effects of a CHANGE intervention to increase exercise maintenance following cardiac events. *Ann Behav Med* 2006; 31: 53–62.
38. Bull FC, Al-Ansari SS, Biddle S, et al. World Health Organization 2020 guidelines on physical activity and sedentary behaviour. *Br J Sports Med* 2020; 54: 1451–1462. 2020/11/27.
39. Shcherbina A, Mattsson CM, 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: 20170524. DOI: 10.3390/jpm7020003.
40. Sartor F, Papini G, Cox LGE, et al. Methodological shortcomings of wrist-worn heart rate monitors validations. *J Med Internet Res* 2018; 20: e10108. 2018/07/04.