Original research article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 11: 1–17

DOI: 10.1177/20552076251337598
journals.sagepub.com/home/dhj

Sage

Continuous heart rate measurements in patients with cardiac disease: Device comparison and development of a novel artefact removal procedure

Paulien Vermunicht[1,2,*], Katsiaryna Makayed[3,4,*], Christophe Buyck[1,2], Lieselotte Knaepen[1,2,5,6], Juan Sebastian Piedrahita Giraldo[3,4], Sebastiaan Naessens[2], Wendy Hens[2], Emeline Van Craenenbroeck[1,2], Kris Laukens[3,4], Lien Desteghe[1,2,5,6] and Hein Heidbuchel[1,2,5]

Abstract

Introduction: Heart rate (HR) monitors could objectively measure physical activity intensity in patients with cardiac disease. However, thorough validation of HR monitors in cardiac populations during daily life, compared to gold-standard Holter monitoring, remains limited. Photoplethysmography (PPG)-based HR data provides near-continuous data, spanning longer periods, but improved algorithms to filter unreliable data are needed.

Methods: This observational, prospective pilot study compared the accuracy of two wearables for HR monitoring (electrocardiogram [ECG]-based Polar H10 chest strap and PPG-based Fitbit Inspire 2 wrist tracker) against Holter monitoring in 15 patients with atrial fibrillation (AF), heart failure (HF) and coronary artery disease referred for cardiac rehabilitation (CR). All devices were worn simultaneously for 24 h. We developed and assessed an artefact removal procedure (ARP) using logistic regression machine learning models to detect unreliable PPG data.

Results: The ECG-based chest strap showed a strong correlation ($r = 0.94$) and clinically acceptable errors (mean absolute error, MAE = 3.4 bpm; mean absolute percentage error, MAPE = 4.9%). Photoplethysmography data exhibited weaker correlation ($r = 0.69$) and higher errors (MAE = 8.3 bpm, MAPE = 14.3%), with highest accuracies in CR and lowest in HF and especially AF. After implementing the ARP, PPG-based HR data improved to a correlation of 0.75, with MAE of 7.2 bpm and MAPE of 12.4%. The procedure removed nearly one-third of unreliable data, achieving an 81% accuracy.

Conclusions: While ECG-based monitors provide HR data with clinical acceptable accuracy, PPG-based monitors present accuracy challenges. Our machine learning procedure showed potential to filter unreliable PPG-based HR data, which could help measure physical activity intensity in cardiac disease continuously.

Keywords

Exercise, cardiac rehabilitation, heart rate, wearable electronic devices, fitness trackers, machine learning

Received: 30 January 2025; accepted: 9 April 2025

[1]Research Group Cardiovascular Diseases, University of Antwerp, Antwerp, Belgium
[2]Department of Cardiology, Antwerp University Hospital, Antwerp, Belgium
[3]Department of Computer Science, University of Antwerp, Antwerp, Belgium
[4]Biomedical Informatics Research Network Antwerp (Biomina), University of Antwerp, Antwerp, Belgium
[5]Faculty of Medicine and Life Sciences, Hasselt University, Hasselt, Belgium
[6]Heart Center Hasselt, Jessa Hospital, Hasselt, Belgium

*Shared first author

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

Introduction

Physical inactivity is closely linked to cardiovascular disease, emphasising the importance of monitoring and enhancing physical activity (PA) in several cardiac populations.[1,2] Existing literature demonstrates the significant role of PA in mitigating the risk and burden of atrial fibrillation (AF),[3–5] in decreasing the risk of developing heart failure (HF) and reducing associated (re)hospitalisations,[6–8] and in improving health-related quality of life for both patients with AF and HF.[9,10] In addition, PA is essential in cardiac rehabilitation (CR) programmes where it contributes to reducing morbidity and mortality and improving cardiovascular risk factors for patients with coronary artery disease.[11]

Heart rate (HR) is a key indicator of PA intensity, reflecting metabolic demand and correlating with oxygen consumption and energy expenditure.[12] This makes it a crucial parameter for personalised exercise prescription.[13,14] As modern healthcare moves more towards home-based care, near-continuous HR monitoring could also be valid to track daily PA in cardiac (tele)rehabilitation programs.[15] Furthermore, HR monitoring could facilitate objective, continuous and non-invasive telemonitoring of PA in a broader population of cardiac patients.

Heart rate monitors exist in various formats. Chest straps, which are electrocardiogram (ECG)-based, are highly accurate but are not suitable for continuous daily monitoring. Wrist-worn devices using the optical photoplethysmography (PPG) technology offer user-friendly alternatives, but PPG accuracy can be impacted by motion and noise artefacts (i.e., under- and overshooting of the real HR).[16,17] While several validation studies have investigated HR monitors in controlled settings, they are mostly limited to short-duration protocols in healthy volunteers. Profound validations in daily-life settings and across different cardiac patient populations are still lacking, particularly studies that compare HR monitors against gold-standard Holter monitoring.[18–20]

Recent advances in machine learning (ML) offer powerful opportunities for processing large and complex datasets, exploring relations between variables and identifying trends and insights from extensive data.[21] These techniques have been widely used in healthcare, enhancing the accuracy of data interpretation, and hence quality of care and patient outcomes.[22] Machine learning is also a promising approach to artefact recognition and removal in PPG signals, but it is important to distinguish between two levels at which these methods can operate. The first level involves ML algorithms that filter artefacts from raw PPG signals, that is, pulsating physiological waveforms representing changes in blood volume due to heartbeats.[23–25] These algorithms are either proprietary, embedded within PPG devices and potentially validated by the manufacturers, but they are not accessible to users or researchers. Alternatively, academic research-based algorithms exist that can only be applied to raw PPG signals, which are typically accessible only through specialised research equipment and have limited validation.[26] At this level, the ability of researchers to improve and validate artefact detection is limited. The second level involves applying ML to the derived HR values obtained from raw PPG signals, obtained as output from commercially available consumer PPG-based HR monitors. Currently, there is a lack of ML algorithms tailored to work directly with these more readily available HR data, with the aim to be used for HR-based assessment of PA load in a medical setting.

The aim of this study was to assess the accuracy of two commercially available HR monitors (Fitbit Inspire 2 wrist-worn fitness tracker and Polar H10 chest strap) compared to gold-standard Holter in patients with AF, HF and following CR after a coronary event or intervention, worn during daily activities. Additionally, this study assessed the potential of ML models to recognise artefacts that impact the accuracy of PPG-HR data, striving for an improved and automatic way to near-continuously track HR in patients with cardiac disease.

Methods

Participants and study design

An observational, prospective, pilot study was designed to explore the accuracy of commercially available HR monitors in patients with cardiac disease and to inform the development of artefact recognition and removal procedures. The study protocol was approved by the Ethics Committee of Antwerp University Hospital and the University of Antwerp (EC reference: 19/21/264, BUN: B300201941069), and all participants provided written informed consent prior to enrolment.

Fifteen participants aged ≥18 years with a scheduled 24-h Holter monitor participated between September 2021 and February 2022. All participants simultaneously wore the Fitbit Inspire 2 wrist-worn fitness tracker and the Polar H10 chest strap along with the gold-standard Holter monitor that served as the reference device. These devices will be further referred to as Fitbit, Polar and Holter. Participants wore the devices simultaneously and continuously in their home environment for 24 h (experimental setup in Supplementary Figure 1). They were advised to maintain their usual routines and asked to record their activities in a diary.

The study comprised three groups of patients with cardiac disease. Inclusion and exclusion criteria were defined per group as follows: *(1) Patients with AF (n = 5)*: inclusion criteria were age ≥18 years and a diagnosis of permanent or persistent AF, with confirmed AF rhythm at the time of Holter monitoring. Patients were excluded if they had a pacemaker or concomitant HF with reduced ejection fraction (HFrEF). *(2) Patients with HF (n = 5)*: inclusion

criteria were age ≥18 years and a diagnosis of HFrEF (left ventricular ejection fraction, LVEF, <40%) with New York Heart Association classification ≥ 2, in a clinically stable condition. Patients were excluded if they had AF. *(3) Patients in CR (n = 5)*: inclusion criteria were age ≥18 years and participation in a hospital-based CR programme following a recent coronary event or intervention (e.g., myocardial infarction, percutaneous coronary intervention or cardiac surgery). Patients with HFrEF or AF were excluded.

The recorded activities varied among participants and included: two supervised training sessions (1 h) by two patients in CR, three home exercise sessions (15–75 min) by two patients in CR and one with AF, and ten cycling sessions (20–90 min) across two patients with AF, two in CR and one with HF.

Heart rate monitors and data collection

The Fitbit utilises optical PPG technology, while the Polar employs electrical field signals to measure HR, both of which are processed through proprietary algorithms exclusive to the respective manufacturers. These devices were selected because they allow export of continuous 24-h HR data, a crucial requirement for this study, which is not supported by many other HR monitors. During data collection, HR data was stored locally on each device, so no active internet or phone connection was required during the recording period. Following the 24-h monitoring period, both Fitbit and Polar data was synchronised by trained study staff to pseudonymised accounts through the Fitbit and Polar Beat smartphone applications, respectively. Subsequently, HR data was exported as comma-separated values (CSV) files using the Polar Flow web service and a Fitbit Web Application Programming Interface. HR data was received at different sampling frequencies: 1-s intervals for Polar and at 5-s intervals for Fitbit. Holter data was exported as CSV files using Sentinel software (Spacelabs Healthcare).

Data processing and alignment

The Holter data, represented as successive beats with timestamps, was converted into HR data as beats per minute (bpm) at 1-s intervals. Extreme values, defined as a 20% difference or more from surrounding HRs, were replaced by their average values. To be comparable with Polar and Fitbit data, Holter data was further processed as a moving average of 3 data points or aggregated into 5-s HR data, respectively. For both the Fitbit and Polar data, extreme outliers beyond the range of physiological plausibility (HR ≤ 25 bpm, HR ≥ 220 bpm) were removed from the data. Next, missing data points were estimated using a polynomial interpolation method, which fits a smooth curve through existing data points when the duration of data absence was less than 30 s. For longer gaps, no interpolation was applied. The proportion of missing data was 39.7% for Fitbit and 0.3% for Polar. For Fitbit, these missing data points predominantly occurred in short gaps due to its default 5-s sampling interval and occasional interruptions. Specifically, 46.3% of the gaps lasted 10 s (one missing HR value), 51.8% lasted 15 s (two missing HR values) and only 1.7% exceeded 15 s. Furthermore, all three devices were perfectly aligned in time by applying data shifts as needed.

Artefact removal procedure

An ML approach was developed to recognise artefacts in Fitbit PPG-HR data and to distinguish these from real activities (based on Holter and diary information). The goal of our procedure was to detect and reject unreliable data sections and only retain trustable data. Such approach was not developed for the Polar data as it already had proven high accuracy in our data and in literature.[27,28] This artefact removal procedure (ARP) includes three steps: (1) Dataset preparation, consisting of feature calculation, episode detection, episode labelling and data aggregation (illustrated in Figures 1–3); (2) Model training for artefact and activity detection; and (3) Combining both models and removing unreliable data.

Step 1: dataset preparation. This involved manually creating a dataset required for training two models to recognise artefacts and activities. First, several features for dynamics analysis were calculated at the 5-s HR level. Second, HR data was split into episodes, which are variable-length segments of the HR time series data determined based on changes in HR dynamics and diary information. Next, target labels for both activity and artefact models were assigned to every episode. Finally, all data was aggregated, transforming each episode into a single data point with its associated features and target labels (Figure 1).

Feature creation: The generated features were designed to capture HR dynamics considering various windows and analysis types. The Savitsky–Golay (SG) filter was employed as a convolutional filter to examine signal waveforms and identify periods of rapid HR changes, such as steep increases and decreases. These changes often correspond to the onset and end of artefacts. Additionally, the Z-score (i.e., a peak detection algorithm implemented in Python) was employed to detect intervals of consecutive signal increases or decreases, disregarding minor fluctuations. This algorithm is used to determine the onset and end points of activities. The scipy.signal package (v1.9.3) was used for the calculations.

Episode detection: The prominence method, implemented in the scipy.signal package, was used to slice the HR data into episodes of varying length, which might represent activities and/or artefacts (Figure 2). This technique isolates independent HR peaks based on their prominence and

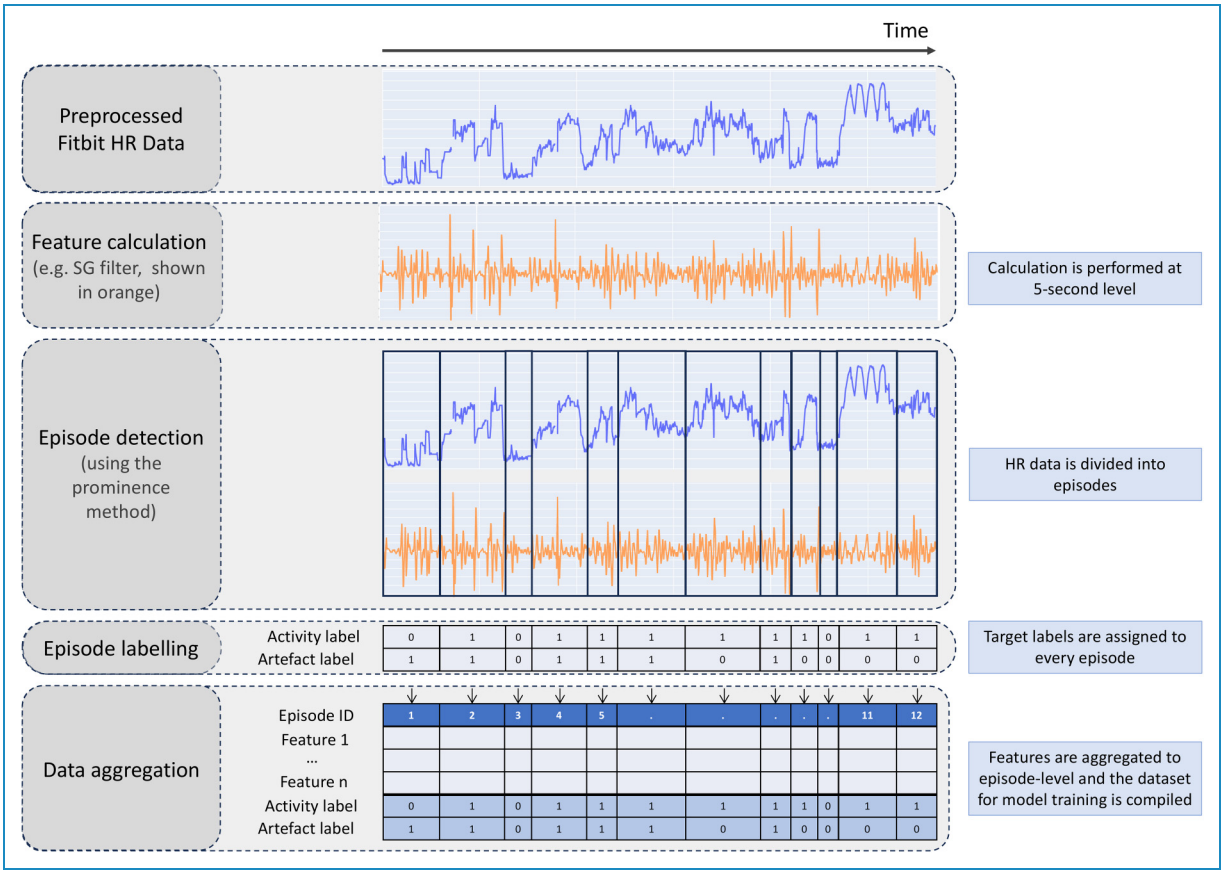

Figure 1. Artefact removal procedure: overview of the manual dataset preparation for model training.

determines the start and end points of episodes. Prominence is defined as the vertical distance in bpm between a peak and its surrounding baseline, which corresponds to the lowest contour line separating the peak from others. Episode detection for Fitbit data followed a two-step procedure. First, the detection algorithm was applied to Holter HR data using diary-reported activities as true labels to optimise a set of four parameters: prominence, width, relative height and rolling window. Episodes with HR peaks meeting the threshold for high prominence (as determined through parameter optimisation) – indicating high HR compared to surrounding baseline – were labelled as 'Holter activity episodes'. Second, the same procedure was applied to detect episodes in Fitbit HR data, using the 'Holter activity episodes' as true labels for parameter optimisation.

Episode labelling: Each episode was attributed two independent target labels – one for activity and another for artefact (Figure 3). Fitbit episodes that aligned with 50% of the Holter activity episodes were labelled as activities, corresponding to real PA detected by the golden standard device. The artefact label was assigned to episodes in which at least 30% of the data deviated from the values recorded by the golden standard device (mean absolute percentage error [MAPE] ≥10%). The 10% MAPE threshold is based on previous literature,[18,29,30] while the 50% activity and the 30% artefact threshold were selected as practical and clinically meaningful by expert consensus within our research team, consisting of clinicians and data scientists.

Data aggregation: HR data at 5-s level was aggregated to the Fitbit HR episodes, applying aggregation functions for the calculated features (e.g., minimum, maximum, mean, standard deviation) and assigning the corresponding target labels. Aggregation functions were applied to the following features: HR data, SG filter values, Z-score values and prominence values (height and length of the episode).

Step 2: model training for artefact and activity episode detection. As episodes may represent both artefacts and activities, two independent logistic regression classification models were trained. The models were trained with 5-fold cross-validation, where 80% of the data were used for training and 20% for validation within each fold. The testing dataset consisted of the full dataset (100%) to evaluate models' performance. Feature selection was performed within each fold using LASSO regularisation, which automatically retained only the most important features, and feature correlation analysis to remove redundant features that provide overlapping information. Additionally, hyperparameter

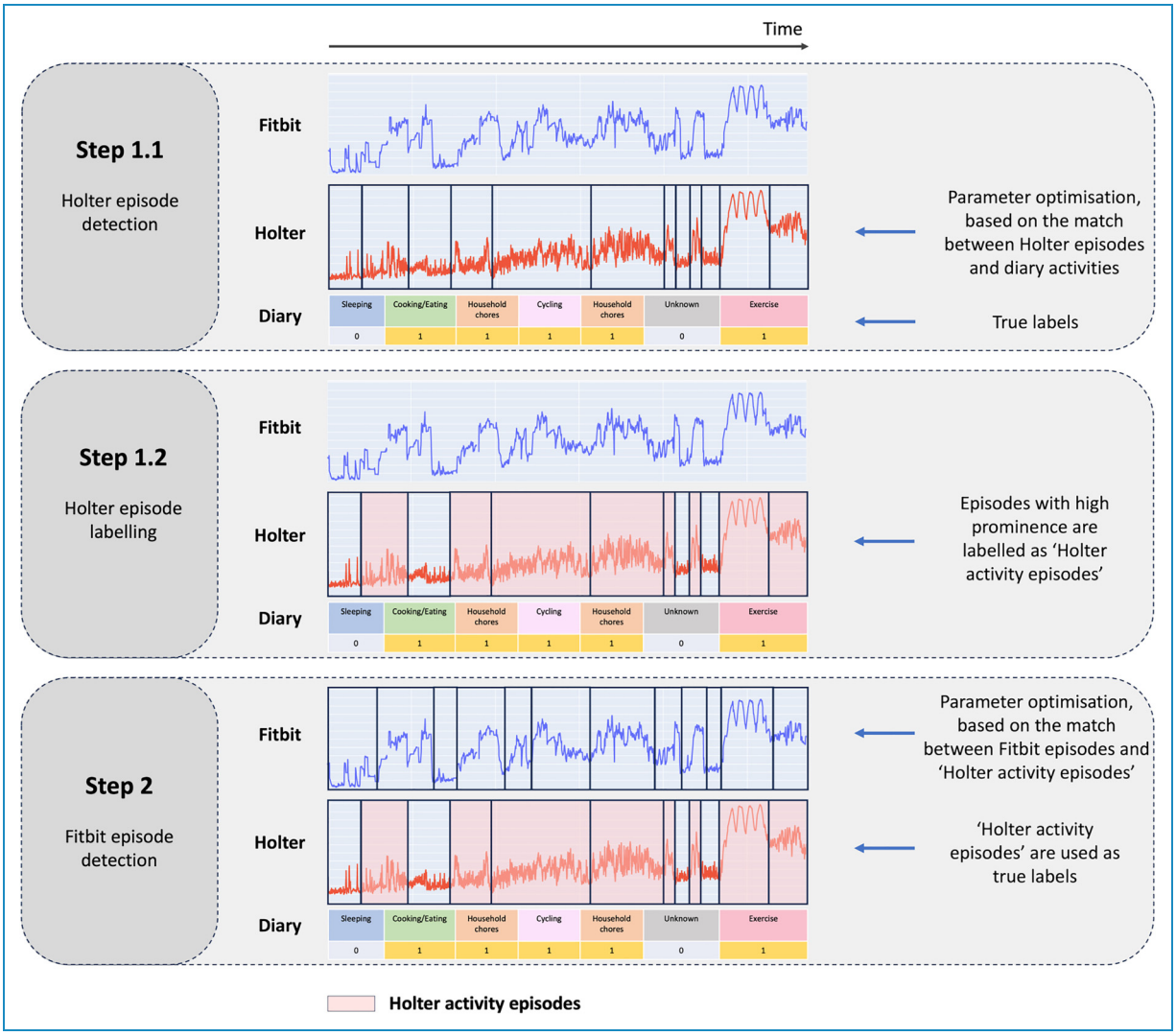

Figure 2. Artefact removal procedure: overview of the episode detection process using the prominence method.

optimisation was conducted by assessing several parameter combinations, with their performance evaluated using cross-validation. This ensured the final models used the best possible settings for optimal performance. The classification models were evaluated separately by calculating following metrics: area under the receiver operating characteristic curve (AUC), accuracy, sensitivity and specificity, with performance metrics averaged across the five folds of the cross-validation process for the validation data.

The computational complexity of the procedure was assessed based on training and application time. Model training using logistic regression, cross-validation and hyperparameter tuning was performed on a standard desktop computer (Intel(R) Core(TM) i5-9500 T CPU @ 2.20 GHz, 8 GB RAM). The training of both the artefact and activity models combined took approximately one hour. Once trained, the models were lightweight and fast to apply to new data: inference on 24 h of HR data for a single user took approximately five seconds, including preprocessing and both model predictions.

Step 3: combining both models and removing unreliable data. The two trained models were applied to the full dataset, resulting in prediction scores at the episode level. The outputs of both models were combined by labelling episodes with high scores of being artefacts and low scores of being activities as 'unreliable', based on optimised thresholds tailored to each model's performance. Labels were merged into the original Fitbit HR data at 5-s level to facilitate the removal of unreliable datapoints. After combining the outputs of both models, the performance of the ARP was evaluated by calculating accuracy, sensitivity and specificity on the testing dataset (100%). The ARP's impact was further assessed by comparing the cleaned data with the original data. A visual example of the decisions of the ARP is provided in Supplementary Figure 4.

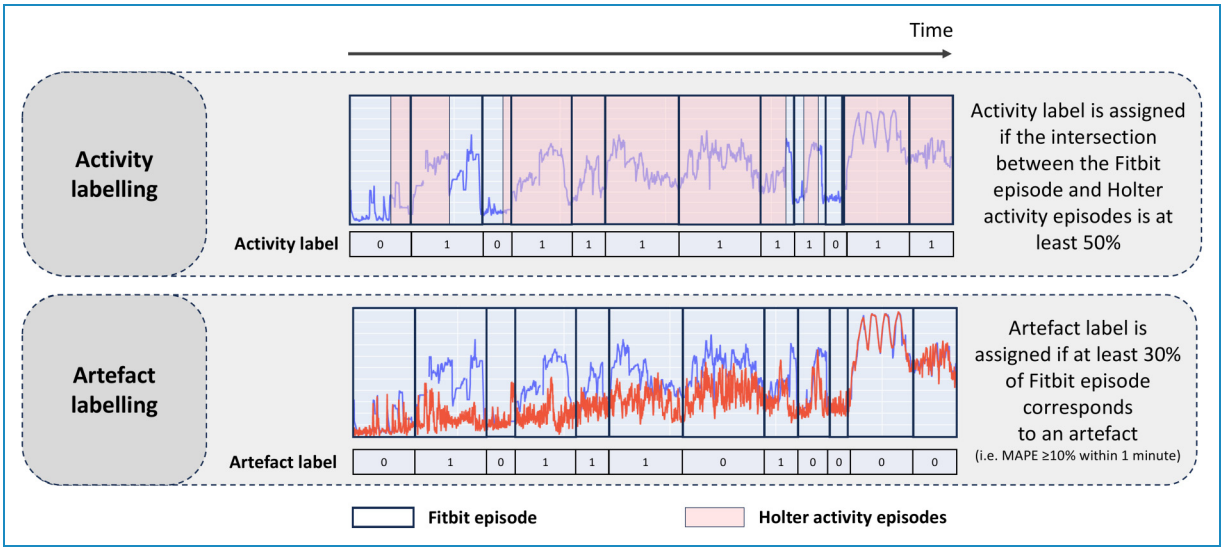

Figure 3. Artefact removal procedure: overview of the episode labelling for activity and artefact classification.

Statistical analysis

All statistical analyses were performed using SPSS Statistics version 29 (IBM Corp) and Python version 3.9. To assess the accuracy of Fitbit and Polar compared to the Holter monitor, various analyses were conducted. The Pearson's correlation coefficient (r) was used to assess linear agreement between HR measurements from the Fitbit and the Holter, and the Polar and the Holter, categorised as negligible ($r = 0.0$–0.30), low positive ($r = 0.30$–0.50), moderate positive ($r = 0.50$–0.70), high positive ($r = 0.70$–0.90) or very high positive correlation ($r = 0.90$–1.00). Next, device error was quantified using mean absolute error (MAE) and MAPE, as commonly applied in validation studies of HR monitors. Generally, an MAPE lower than 10% is considered acceptable based on previous research and on the standard for HR monitors by the American National Standards Institute.[17,28,29] This 10% threshold was also used for labelling artefacts. The percentage difference, without taking absolute values, was calculated to differentiate between undershooting (percentage error $\leq -10\%$) and overshooting (percentage error $\geq 10\%$). In a Bland–Altman analysis, the mean bias and 95% CI limits of agreement (LoA) were calculated and presented graphically to estimate any tendency for variation to change with the magnitude of HR. Additionally, the MAPE was used to categorise patients into PPG compatible (MAPE $< 10\%$) and PPG incompatible (MAPE $\geq 10\%$) groups, in order to account for potential patient-specific inaccuracies influencing overall group trends and the performance of the ARP.

All p-values were two-sided, and a significance level of $p < 0.05$ was used throughout. Normality of continuous variables was assessed using the Shapiro–Wilk test and visual inspection of histograms. When normality could not be assumed, non-parametric tests were selected accordingly. Group-level comparisons (AF, HF, CR) for demographic and clinical baseline characteristics were performed using Kruskal–Wallis tests for non-paired continuous variables (e.g., age, weight) and Fisher's exact test for categorical variables (e.g., gender, skin type), appropriate for small cell sizes. To compare mean HR values between devices (Holter, Fitbit, Polar), the Friedman test was applied for continuous paired variables. When significant, post hoc pairwise comparisons were performed using the Wilcoxon matched-pairs signed-rank test. Cohen's d was calculated to determine effect sizes for significant differences, using established thresholds for interpretation (e.g., small: 0.2, moderate: 0.5, large: 0.8).

Results

Demographics

Baseline characteristics of the 15 participants are summarised in Table 1. Predominantly men (87%) participated in the study, with a median age of 62.0 (interquartile range: 52.0–75.0) and a median weight of 81.6 kg (interquartile range: 69.0–93.0). Most participants (80%) had light skin colour. In the AF group, 100% of the recording time was characterised by AF, while in the HF and CR groups, AF was observed 0% of the time.

Accuracy of Fitbit Inspire 2 and Polar H10 before artefact removal

Overall, 1,267,255 data points were collected with the Holter monitor, 253,616 with the Fitbit device, and 1,251,366 with the Polar monitor, reflecting the devices'

Table 1. Baseline characteristics of the study population.

	Total ($n = 15$)	AF group ($n = 5$)	HF group ($n = 5$)	CR group ($n = 5$)	p value
Demographic and clinical characteristics					
Male, *n* (%)	13 (86.7)	5 (100.0)	3 (60.0)	5 (100.0)	0.29
Age (years), median (IQR)	62.0 (52.0–75.0)	75.0 (64.0–79.0)	63.0 (55.0–75.5)	52.0 (44.5–60.0)	**0.04**
Weight (kg), median (IQR)	81.6 (69.0–93.0)	90.0 (80.5–122.5)	69.0 (60.5–91.0)	78.0 (70.5–87.3)	0.14
BMI (kg/m^2), median (IQR)	25.0 (23.3–29.7)	29.7 (26.1–42.6)	25.0 (22.1–30.6)	24.4 (23.0–26.9)	0.15
Skin type					
Light skin, n (%)	12 (80.0)	5 (100.0)	4 (80.0)	3 (60.0)	0.73
Lightly toned skin, n (%)	2 (13.3)	0 (0)	1 (20.0)	1 (20.0)	1.00
Asian type, n (%)	1 (6.7)	0 (0)	0 (0)	1 (20.0)	1.00
Mean eGFR (mL/min/1.73m²)	71.3	73.0	61.4	77.4	0.06
eGFR ≤50 mL/min/1.73m², *n* (%)	2 (13.3)	0 (0)	2 (40)	0 (0)	0.45
Signs of congestion, *n* (%)	3 (20.0)	1 (20.0)	2 (40.0)	0 (0)	1.00
Cardiovascular risk factors					
Diabetes mellitus, *n* (%)	2 (13.3)	1 (20.0)	0 (0)	1 (20.0)	1.00
Hypertension, *n* (%)	9 (60.0)	4 (80.0)	3 (60.0)	2 (40.0)	0.80
CVA/TIA, *n* (%)	0 (0)	0 (0)	0 (0)	0 (0)	NA
Vascular disease, *n* (%)	11 (73.3)	2 (40.0)	4 (80.0)	5 (100.0)	0.23
Prior MI, *n* (%)	6 (40.0)	0 (0)	2 (40.0)	4 (80.0)	0.07
HFrEF, *n* (%)	6 (40.0)	0 (0)	5 (100.0)	1 (20.0)	**0.006**
Hypercholesterolemia, *n* (%)	12 (80.0)	4 (80.0)	4 (80.0)	4 (80.0)	1.00
Obesity, *n* (%)	4 (26.7)	2 (40.0)	2 (40.0)	0 (0)	0.45
Smoking, *n* (%)	0 (0)	0 (0)	0 (0)	0 (0)	NA

AF: atrial fibrillation; HF: heart failure; CR: cardiac rehabilitation; IQR: interquartile range; BMI: body mass index; CVA: cerebrovascular accident; TIA: transient ischemic attack; MI: myocardial infarction; HFrEF: heart failure with a reduced ejection fraction (left ventricular ejection fraction, LVEF, <40%); eGFR: estimated glomerular filtration rate; NA: not applicable. Fluid status based on clinical signs (e.g., edema, orthopnea, jugular vein distension) and/or echocardiographic signs of elevated filling pressures. *P* values in bold are statistically significant ($p < 0.05$).

respective sampling frequencies as described in 'HR monitors and data collection' section. The mean HR was 67.3 ± 18.0, 71.7 ± 16.1 and 67.5 ± 17.5 with the Holter, Fitbit and Polar monitors, respectively, with significant differences: Holter versus Fitbit ($p < 0.001$, Cohen's $d = 0.32$) showing a small to moderate effect size, and Holter versus Polar ($p < 0.001$, Cohen's $d = 0.03$) showing a very small effect size.

Correlation and device error analysis. Polar data exhibited a very high positive correlation ($r = 0.94$) with the Holter data, with an MAE of 3.4 bpm and an MAPE of 4.9%. The MAPE remained below the 10% criterion across all patient groups for Polar. Conversely, Fitbit Inspire 2 HR data displayed a moderate positive correlation ($r = 0.69$), with an MAE of 8.3 bpm and an MAPE of 14.3%. These observations varied across patient groups (Table 2), with

Table 2. Correlation and device error analysis, before artefact removal.

	Correlation analysis	Device error	
	Pearson correlation coefficient (r)	MAE (bpm), mean ± SD	MAPE (%), mean ± SD
Fitbit Inspire 2 fitness tracker			
All patients (n = 15)	0.69	8.3 ± 5.5	14.3 ± 13.6
AF (n = 5)	0.67	12.1 ± 5.6	21.1 ± 18.3
HF (n = 5)	0.57	6.8 ± 5.7	11.7 ± 12.9
CR (n = 5)	0.79	5.9 ± 3.9	10.0 ± 7.5
Polar H10 chest strap			
All patients (n = 15)	0.94	3.4 ± 2.6	4.9 ± 3.6
AF (n = 5)	0.92	6.4 ± 1.8	9.1 ± 1.4
HF (n = 5)	0.93	2.4 ± 2.0	3.4 ± 2.7
CR (n = 5)	0.99	1.5 ± 0.8	2.2 ± 1.3

MAE: mean absolute error; bpm: beats per minute; SD: standard deviation; MAPE: mean absolute percentage error (criterion: ≤10%); AF: atrial fibrillation; HF: heart failure; CR: cardiac rehabilitation. The strength of the correlation was interpreted as negligible (r = 0.0–0.30), low positive (r = 0.30–0.50), moderate positive (r = 0.50–0.70), high positive (r = 0.70–0.90) or very high positive (r = 0.90–1.00).

the highest accuracies observed in the CR group and the lowest in the AF group (based on MAE/MAPE) and in the HF group (based on correlation).

In Figure 4, the type of artefacts was analysed by distinguishing between undershooting (percentage difference ≤−10%) and overshooting (percentage difference ≥10%) compared to the Holter HR. The Fitbit device exhibited a higher percentage of overshooting (25.7%) compared to undershooting (7.5%), a trend consistent across all participant groups, with the highest overall percentage of under/overshooting observed in the AF group (54.6% compared to 24.0% in HF and 20.9% in CR). In contrast, the Polar monitor demonstrated less overall under/overshooting (15.4% compared to 33.2% in Fitbit). Similarly, in the AF group, Polar showed the highest percentage of under/overshooting (35.4% compared to 7.8% in HF and 2.8% in CR).

Day and night differences. Based on the patients' self-reported sleep times, 34.8% of all data was classified as night time. Fitbit showed moderate positive agreement (r = 0.58) with the Holter monitor during the day, improving to high positive agreement (r = 0.80) at night (Table 3, Supplementary Figure 2). Additionally, MAE and MAPE decreased by 4.9 bpm and 6.0%, respectively, and the percentage of under/overshooting was reduced by 14.6% during the night. The Polar device demonstrated a high positive correlation (r > 0.85) both during the day and night, with similar patterns in other accuracy metrics. When distinguishing between the groups, improvements in accuracy during the night were smallest in the AF group, or sometimes even absent (Supplementary Table 1).

Bland–Altman analysis. Bland–Altman analysis (Figure 5) on the 24-h data showed that Fitbit generally tended to overestimate HR by a mean bias of 4.3 bpm, especially at lower HRs, with more underestimation observed as HR increased, while the Polar showed minimal bias (0.1 bpm). The LoA's for Fitbit (−22.4 to 31.0 bpm) were wider than those for Polar (−12.7 to 12.9 bpm). Bland–Altman plots, presented separately for each group, revealed that in the AF and HF group, Fitbit tended to overestimate lower HRs (<±60 bpm), while higher HRs (>±90 bpm) were more likely to be underestimated. In the CR group, Fitbit overestimated lower HRs (<±60 bpm), but except for a few outliers, there were no significant under- or overestimations when HRs were higher than ±100 bpm. Colour-coding for individual patients demonstrates that patient-specific tendencies in different directions can be prominent, particularly in the HF and CR groups. For instance, HR data of patient 18 in the HF group, who had a notably higher percentage of ventricular ectopic beats (18%) compared with 0–2% in the other patients, showed an underestimation that deviated from the group tendency. In the CR group, Polar data of patient 7 showed an off-trend underestimation, occurring before and after a period of connectivity loss of the Polar device.

Bland–Altman analyses were also performed on daytime data only (Supplementary Figure 3) given its relevance for PA monitoring, but since the conclusions were unchanged, the 24-h data are presented throughout the paper.

Performance of the ARP in Fitbit inspire 2 data

The total dataset consisted of 253,616 data points divided into 636 HR episodes (average duration: 33m14 s ± 57m36 s and range: 15 s – 9h40m10 s). A total of 65,525 datapoints were labelled as true artefacts and an additional 99,233 as true activities. The artefact model achieved 80% ± 1% accuracy, 92% ± 1% sensitivity, 59% ± 2% specificity and an AUC-value of 84% ± 2% based on the validation data from the 5-fold cross-validation process, while the activity model reached 81% ± 1% accuracy, 81% ± 1%

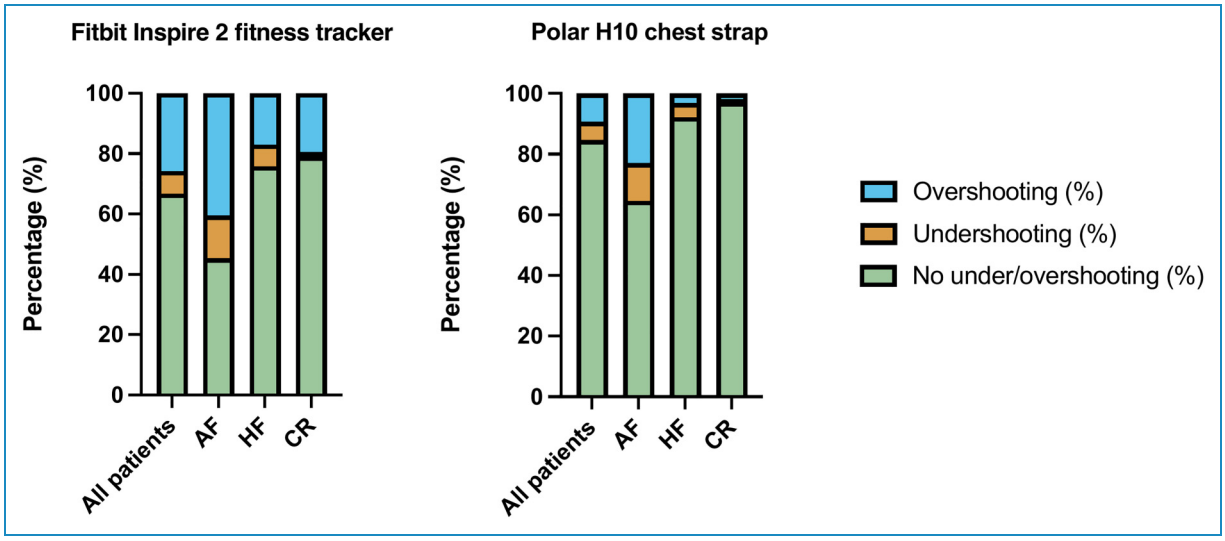

Figure 4. Percentages of under- and overshooting, before artefact removal. AF: atrial fibrillation; HF: heart failure; CR: cardiac rehabilitation. Under- and overshooting were defined as a percentage difference of 10% or more compared to the Holter device.

Table 3. Day versus night accuracy analysis for all patients ($n = 15$), before artefact removal.

	Fitbit Inspire 2 fitness tracker		Polar H10 chest strap	
	Day	Night	Day	Night
Pearson correlation coefficient (r)	0.58	0.80	0.94	0.93
MAE (bpm), mean ± SD	10.0 ± 6.6	5.2 ± 5.6	3.5 ± 2.7	3.6 ± 3.2
MAPE (%), mean ± SD	16.4 ± 14.3	10.3 ± 15.8	4.6 ± 3.4	5.6 ± 4.2
Under/overshooting, total (%)	38.4	23.8	14.2	17.6
Undershooting (%)	8.8	5.8	6.2	5.9
Overshooting (%)	29.5	18.0	8.0	11.7

MAE: mean absolute error; bpm: beats per minute; SD: standard deviation; MAPE: mean absolute percentage error. The strength of the correlation was interpreted as negligible ($r = 0.0$–0.30), low positive ($r = 0.30$–0.50), moderate positive ($r = 0.50$–0.70), high positive ($r = 0.70$–0.90) or very high positive ($r = 0.90$–1.00). Under- and overshooting was defined as a percentage difference of 10% or more compared to the Holter device. Day and night times were personalised based on patients' self-reported sleeping time, except for one participant for whom no information was available, and standard timing was used instead (sleep from 12:00 am to 6:00 am).

sensitivity, 82% ± 5% specificity and an AUC-value of 87% ± 2% using the same validation approach.

Combining the results from both models, the testing dataset (100% of the data) was used to assess the procedure's overall performance. Of the complete dataset, 19,994 data points (7.9%) were labelled as unreliable and were removed by our procedure. Of these, 18,757 were true artefacts (correctly labelled true positives) and 1237 were incorrectly labelled (false positives). The procedure labelled 233,622 HRs as trustable data (92.1%), correctly identifying 186,854 (true negatives) while incorrectly labelling 46,768 artefacts as reliable (false negatives). This resulted in an overall accuracy of 81%, sensitivity of 28% and specificity of 97%.

For illustrative purposes, Supplementary Figure 4 provides a visual representation of the ARP using example data of a patient following CR, where our procedure identified multiple artefacts (though not all), without removing activities from the data. Out of the 99,233 true activity HRs, the procedure correctly removed 10,882 as true under/overshootings (11.0%). However, 1169 activities were mistakenly removed despite not being under/overshootings (1.2%). The procedure correctly retained 65,741 (66.2%) activities as trustable HR measurements, but incorrectly retained 21,441 (21.6%) activities that were still under/overshootings of the HR.

The impact of the ARP on validity metrics, including the correlation coefficient, MAE, MAPE and the total percentage of under- and overshooting, is depicted as absolute values in Figure 6 and as percentage improvements in Supplementary Figure 5. Across all groups, excluding unreliable data resulted in enhancements in all validity metrics: correlation from 0.69 to 0.75, MAE from 8.3 to 7.2 bpm, MAPE from 14.3% to 12.4% and under/overshooting from 33.2% to 29.7%. The smallest gains were observed in the AF group. To address potential patient-specific inaccuracies in PPG-HR measurements, patients were categorised as exhibiting PPG compatibility (MAPE < 10%, $n = 7$) or PPG incompatibility (MAPE ≥ 10%, $n = 8$), as

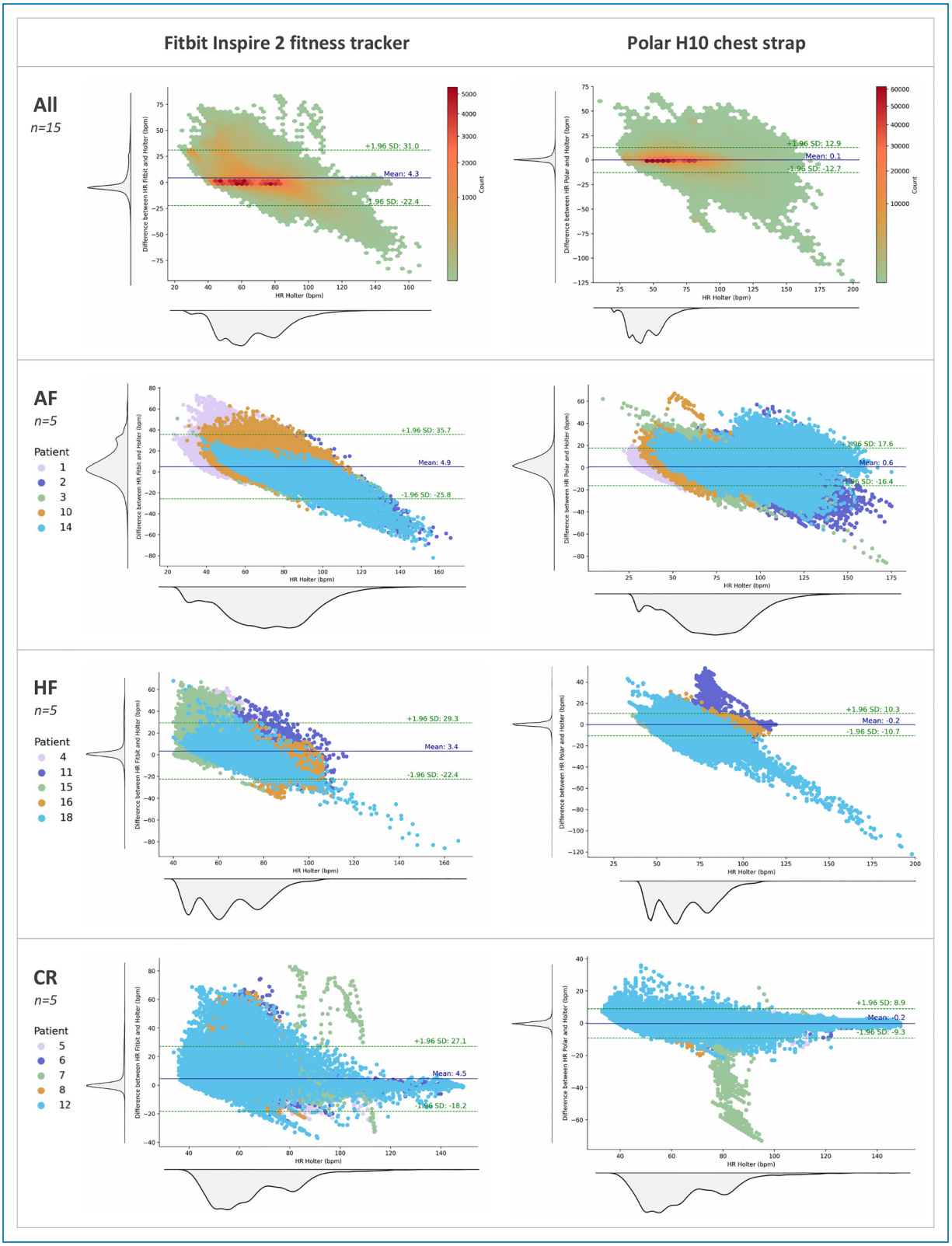

Figure 5. Bland–Altman density plots for 24-h data before artefact removal: comparison between all patients (all) and the atrial fibrillation (AF), heart failure (HF) and cardiac rehabilitation (CR) groups, colour-coded for individual patients. AF: atrial fibrillation; HF: heart failure; CR: cardiac rehabilitation; HR: heart rate; SD: standard deviation. A positive mean (bias) indicates that the Fitbit/Polar overestimates HR compared to the Holter, a negative mean (bias) indicates that the Fitbit/Polar underestimates HR; limits of agreement (LoA, green dotted lines) indicate the range in which 95% of all differences between the two methods lie. Individual patient data are colour coded.

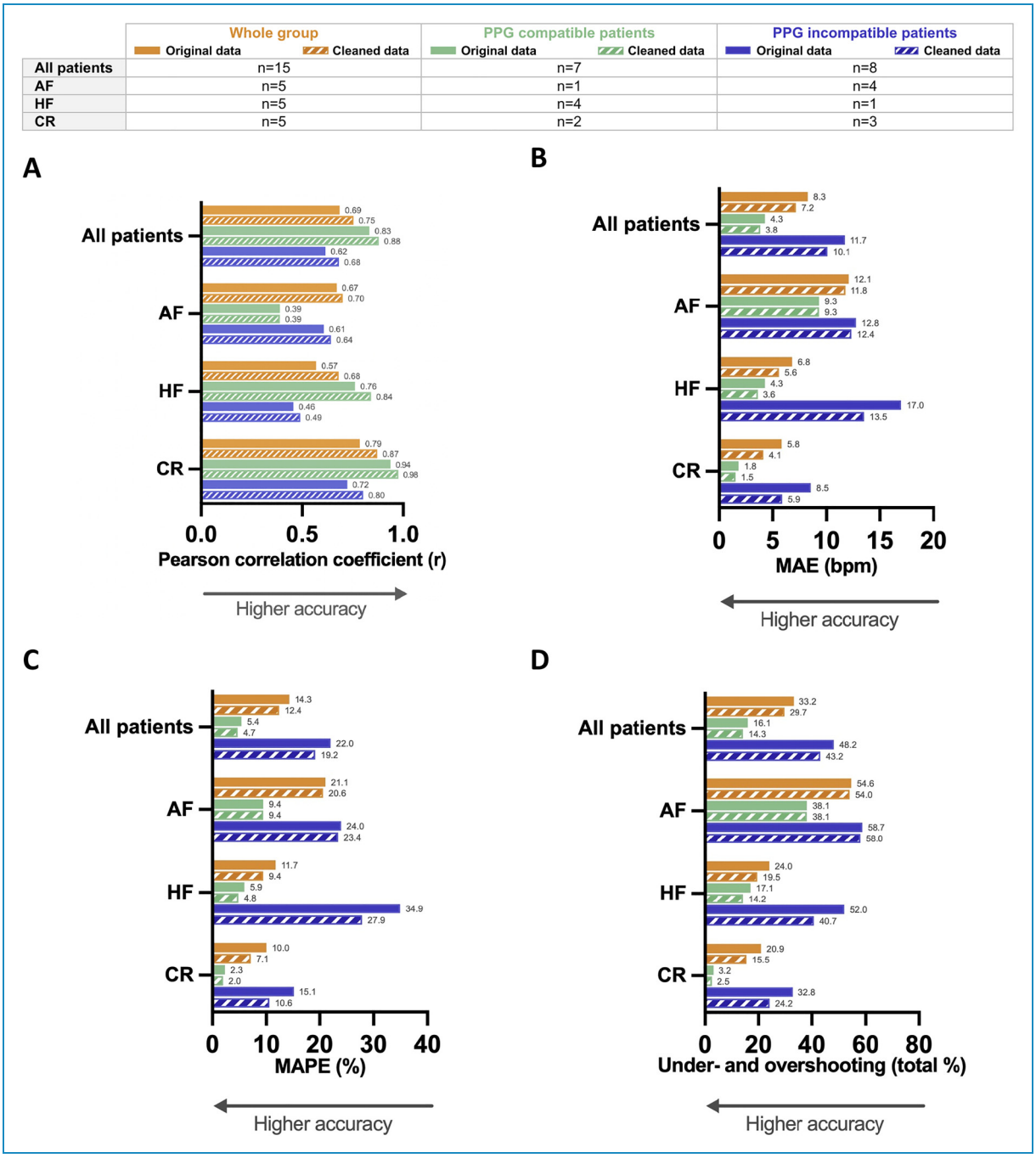

Figure 6. Absolute values of the Pearson correlation coefficient (A), MAE (B), MAPE (C) and total percentage of under- and overshooting (D) before and after removal of unreliable Fitbit data: original versus cleaned data. MAE: mean absolute error; bpm: beats per minute; MAPE: mean absolute percentage error; AF: atrial fibrillation; HF: heart failure; CR: cardiac rehabilitation. Under- and overshooting was defined as a percentage difference of 10% or more compared to the Holter device. Patients were categorised into PPG-compatible (MAPE < 10%) and PPG-incompatible (MAPE ≥ 10%) groups. Cleaned data refers to the data remaining after the artefact removal procedure, with unreliable data removed and only trustable data retained.

described in the Methods (Statistical analysis) section. Accuracy improvements after removing unreliable data with the ARP algorithm were more pronounced in the PGG incompatible group (correlation from 0.62 to 0.68, MAE from 11.7 to 10.1 bpm, MAPE from 22.0% to 19.2% and under/overshooting from 48.2% to 43.2%) and were less pronounced but still relevant in the PPG compatible group. However, the validity in the PPG incompatible group remained insufficient based on the 10% MAPE criterion.

Discussion

Our study is the first to evaluate the accuracy of continuous measurements of two commercially available HR monitors, the PPG-based Fitbit Inspire 2 wrist-worn fitness tracker and the ECG-based Polar H10 chest strap, during a 24-h monitoring period in a clinical cohort of patients with cardiac disease. Since cardiac-related pathophysiological factors such as peripheral edema, poor tissue perfusion and a higher incidence of arrhythmias may influence the PPG signal,[31] results obtained from healthy individuals may be unsuitable for applications targeting patients with cardiac disease.[32] It is therefore important to validate PPG-based HR monitors specifically in cardiac populations, which was the goal of our study. Additionally, we explored ML models to recognise and remove artefacts from HR data obtained from PPG-based HR monitors to improve HR-based PA load prediction.

Confirmed accuracy of ECG-based chest strap

The *Polar H10 chest strap* demonstrated very high correlation ($r = 0.94$) and minimal error (MAE = 3.4 bpm, MAPE = 4.9%) compared to the gold-standard Holter monitor across all patient groups, both at night and during daily activities. Previous studies had primarily evaluated the Polar H10 in healthy volunteers and only during specific exercise protocols. Schaffarczyk et al. and Merrigan et al. reported, respectively, a Pearson correlation of 1.00 and an MAPE of 1.3–3.4% in healthy volunteers during exercise protocols,[33,34] while Etiwy et al. reported a Lin's concordance correlation coefficient (CCC) of 0.99 for the earlier version Polar H7 in cardiac patients during a CR session.[19] Our study is unique in its evaluation of 24-h continuous recordings in specific cardiac patient groups. It confirmed the accuracy and reliability of the Polar H10 chest strap across these cardiac patient populations and outside controlled testing conditions.

Challenges in accuracy of PPG-based wrist-worn HR monitors across cardiac patient populations

The *Fitbit Inspire 2* exhibited a moderate positive correlation ($r = 0.69$) and higher error rates (MAE = 8.3 bpm, MAPE = 14.3%) compared to the ECG-based chest strap. The PPG-based device showed more overshooting (25.7%) than undershooting (7.5%), often overestimating HRs at lower levels and underestimating them above 100 bpm. Accuracy was higher at night, likely due to less movement, consistent with prior findings.[35] The poorer accuracy compared to the chest strap was expected, given the limitations of optical PPG technology, such as sensitivity to motion artefacts and technical factors such as improper fitting or skin contact.[36]

Comparisons between PPG-based devices, such as Apple Watch, Garmin and Polar, suggest that device-specific performance may vary, underscoring the need for tailored validation of each device for clinical use.[19,37] Although the Fitbit Inspire 2 HR monitor itself has not yet been validated in the literature, analogous Fitbit devices have been studied, and these data will be used as a basis for comparison in this discussion. Regarding studies in *healthy volunteers*, Nelson et al. reported lower error rates for the Fitbit Charge 2 (MAE = 3.5 bpm, MAPE = 6.0%) compared to our findings (MAE = 8.3 bpm, MAPE = 14.3%) during a 24-h monitoring period with ECG validation. However, their study only included one participant.[18] Another study found a CCC of 0.83 for the Fitbit Charge HR tracker in 10 healthy volunteers during activities, but this comparison used a chest strap rather than Holter, and the measurements were assessed in 1-min intervals rather than continuously.[38] Additionally, various studies validated Fitbit devices in healthy volunteers during treadmill or bike sessions, but these results vary widely (CCC: 0.50–0.89, MAPE: 2.38–16.99%) and are hard to compare to our 24-h monitoring period due to the controlled settings.[16,17,27,39–43]

Also, validation of PPG-based HR in *patients with cardiac disease* is limited in prior research. In *patients following CR*, the Fitbit Blaze demonstrated a correlation of 0.78 during an exercise session,[19] aligning with the correlation of 0.79 found in our CR group during 24-h continuous monitoring.

For *patients with AF*, our study's results ($r = 0.67$, MAE = 12.1 bpm) were consistent with previous findings. Quinn et al. reported lower correlations (rest: $r = 0.50$, peak exercise: $r = 0.30$) and higher errors (rest: MAE = 7.0 bpm, peak exercise: MAE = 28.7 bpm) in patients with AF compared to those in sinus rhythm (rest: $r = 0.91$, peak exercise: $r = 0.73$; rest: MAE = 4.6 bpm, peak exercise: MAE = 13.8 bpm).[37] Another study by Al-Kaisey et al. demonstrated a similar correlation ($r = 0.60$) using a Fitbit Charge HR device during 24-h monitoring.[35] Both studies and our findings indicate lower PPG accuracy in patients with AF compared to those in sinus rhythm. The irregular heart rhythm and beat-to-beat variability in cardiac output imply that not every heartbeat generates a sufficiently strong pulse, leading to a phenomenon known as pulse deficit, which affects the amplitude of the PPG signal and contributes to the accuracy challenges for AF rhythms.[35,44] Our patients with AF were also significantly older than the CR group ($p = 0.02$). Ageing contributes to arterial stiffness and causes skin changes such as thinning, wrinkles and hyperpigmentation, all of which affect light interaction with blood vessels and reduce PPG signal quality.[29] Although BMI was higher in the AF group compared to other study populations, it did not reach statistical significance ($p = 0.15$), but could still contribute to lower PPG accuracy due to its impact on skin thickness and blood flow dynamics.[36]

No validation studies are available in the literature for *patients with HF*. In our study, this group showed a weak correlation ($r = 0.57$, inferior to AF and CR groups) and moderate error (MAE = 6.8 bpm, MAPE = 11.7%, worse than the CR group and better than the AF group). The pathology of HF likely contributes to this lower PPG accuracy due to reduced peripheral circulation, fluid retention and vasoconstriction, which affect light absorption and reflection in the skin.[45] Although not statistically significant, our baseline data suggested a trend towards lower renal function and more signs of congestion in the HF group, potentially contributing to the reduced PPG performance.

The observed artefacts and inaccuracies may be clinically important if PPG-based HR monitors were to be used for monitoring PA or prescribing exercise with specific HR targets. Artefacts in the data from these monitors may lead to under/overshooting the amount of PA, false assurance of good HR control or incorrect medication titration.[37] In our study, PPG-based monitoring showed the highest accuracy in the CR group, where MAPE values approached the acceptable 10% threshold, suggesting potential clinical utility in this population. In contrast, patients with AF and HF showed poorer accuracy, indicating that clinical use of PPG-based HR monitoring for PA assessment in these populations remains premature at this stage. To address artefacts and minimise their impact on PA assessments, our research group explored two approaches: (1) an objective preselection of patients based on PPG compatibility (MAPE < 10%), and (2) an ML-based ARP for detecting and eliminating unreliable PPG-HR data. These approaches will be further discussed in 'Necessity of patient preselection for PPG-HR monitoring in clinical settings' and 'Opportunities of the current ARP' sections. While we acknowledge that PPG-based HR measurements in patients with AF and HF may not be suitable for clinical diagnoses or treatment decisions (e.g., bradycardia detection or beta-blocker titration), they may still provide a reasonable estimate of daily PA intensity – particularly during daytime periods when activity occurs, provided that appropriate patient preselection and artefact removal strategies are applied.

Necessity of patient preselection for PPG-HR monitoring in clinical settings

The findings of our study suggest that PPG-based HR monitoring may not be suitable for all patients, even after artefact detection and removal using ML techniques. Specifically, more than half of the patients (53%) in our study showed low overall accuracy (MAPE ≥10%) in their PPG-HR measurements, likely due to patient-related factors impacting PPG reliability. In addition, the ARP was insufficiently effective in detecting artefacts and thus improving accuracy in this group.

Therefore, we now suggest that, before PPG-HR monitoring can be considered for clinical use, an objective preselection of patients may be necessary to determine their suitability for this technology. Patients could be selected based on their initial PPG accuracy (e.g., MAPE < 10%), which would help minimise the risks associated with inaccurate recorded HR data. For the selected patients, the ARP could be applied to remove remaining artefacts and further improve data accuracy. Further research could refine this preselection criterion for routine clinical assessments of PPG compatibility.

Opportunities of the current ARP

To our knowledge, our study is the first to develop a procedure that uses ML techniques to identify artefacts in continuous HR data obtained from commercially available PPG-based HR monitors. Most existing artefact removal methods require access to raw PPG waveforms, which are either embedded in proprietary algorithms or available only through research-grade equipment. In contrast, our approach operates directly on derived HR data – the only data accessible from most consumer devices – making it feasible for real-world use. This approach bridges the gap between academic research and clinical applicability and supports long-term HR monitoring in outpatient settings.

The ARP approach also aligns with current recommendations in the field of medical ML, which advocate for explainable, resource-efficient models that can be applied in real-world clinical environments.[46] Unlike deep learning methods, which are often complex and opaque, our use of logistic regression enables transparent, fast and supervised artefact detection suitable for outpatient settings.[47] This makes our approach suitable for integration into future clinical workflows and mobile health solutions.

The current version of the ARP yielded mixed results. Of all artefacts present, one-third was detected by the procedure, thereby reducing the number of under/overshooting in the HR data. Only 0.50% of the total data was incorrectly labelled as unreliable, indicating minimal erroneous removal of HR data. However, while only 1% of all physical activities in our dataset were incorrectly labelled as unreliable, 21% of activities were considered reliable despite being over- or underestimated conform our current definition (MAPE≥10%). This could still pose challenges if PPG-based HR monitors were used for PA follow-up, potentially leading to an under/overestimation of performed PA.

Notably, the ARP was found to be insufficiently effective in patients with AF, likely due to the irregular heartbeat characteristics typical of this group, making it challenging to distinguish from artefacts, as was already discussed in a prior section.

While the overall results of the procedure appear promising, further training, refinement and validation using new

data are necessary to enhance accuracy and reliability. This will enable the procedure to better distinguish between activities and artefactual high HRs and will hopefully lead to improved usefulness across diverse patient profiles.

Limitations

Many noise sources can affect PPG signals, including individual patient variations (e.g., skin tone, BMI, age, gender) and external or environmental factors (e.g., strap tightness, ambient light, temperature).[36] We could not report or analyse all factors. The same investigators applied the wrist-worn monitor aiming for uniformity (i.e., as tight as possible yet still comfortable), but strap tightness was subjectively determined without objective measurement of tension or pressure. Consequently, device tightness may have varied among participants. Additionally, sensor placement, temperature and ambient light intensity were not reported, although these factors can impact the PPG signal.[32] The data collection period spanned varying ambient temperatures, which may have influenced peripheral circulation and thus affected PPG signal quality. Future studies may consider collecting environmental data or differentiating indoor versus outdoor activities to explore this further. Skin colour and gender were reported in our study, but the variation within our relatively small patient population was insufficiently balanced to perform detailed analyses on differences in PPG accuracy. Previous studies have shown that darker skin leads to reduced PPG signal quality due to the higher concentration of melanin,[36] and that men show lower PPG accuracy compared to women due to physiological differences in skin thickness and arterial stiffness.[42] Our study population was predominantly male (87%) which is a common limitation in cardiovascular research and may have introduced a potential gender bias.[48] Future studies should aim to recruit a more balanced cohort, in terms of both skin colour and gender, to better understand potential patient-related differences in PPG accuracy and their implications for clinical practice.

Another limitation of this study is its small sample size ($n = 15$), divided across three groups ($n = 5$ per group), which may have reduced statistical power and generalisability. However, the use of high-resolution continuous HR data over 24 h per participant provided a sufficiently rich dataset to explore device accuracy. Though our study included three clinically relevant subgroups (AF, HF and CR), the findings may not be generalisable to other populations or cardiac conditions, such as patients with pacemakers or with preserved ejection fraction. Finally, we acknowledge that the internal signal processing of the Fitbit PPG device remains a black box. Specifically, the algorithm's handling of arrhythmias such as ventricular ectopic beats is unknown, and this may have contributed to some of the observed inaccuracies. Nevertheless, this study was intentionally designed as a pilot to provide preliminary insights into the accuracy of HR monitors in cardiac populations and to support the development of the ARP. To address the limitations, the follow-up ARTEPHYISCAL study (NCT05901038) is currently underway and aims to provide a larger and more diverse dataset for more robust analyses and generalisable conclusions.

The current ARP also has limitations as it was developed using data from this small pilot study without the use of an external, unseen testing dataset. This methodological choice was necessary to maximise the use of the available data in this initial development phase, but overfitting, where the model learns patterns specific to the training data rather than generalisable trends, may have inflated the reported performance metrics. Moreover, the dataset included a limited variety of physical activities, complicating differentiation between activities and artefacts. The target activity labelling was partially based on self-reported diary entries, in combination with HR data from the golden standard Holter monitor. This reliance on self-reported data may introduce bias due to potential inaccuracies in patient reporting. Our current classification models use logistic regression to detect artefacts and activities due to its simplicity and interpretability, but this technique may struggle with complex, non-linear relationships, especially in imbalanced datasets like ours. Alternative tree-based methods may offer improved performance in future work. Due to the proprietary nature of PPG signal processing in consumer devices, raw waveform data was not accessible. As a result, our artefact detection procedure was applied to derived HR values, which may limit certain signal-level corrections. However, this reflects the type of data typically available in clinical and real-world applications, supporting the practical relevance of our approach. Further refinement and validation of the procedure is being pursued in the ongoing ARTEPHYISCAL study (NCT05901038), which is collecting a more extensive dataset with standardised physical exertions to address the mentioned challenges, including testing the procedure on independent datasets to ensure generalisability to unseen data.

Conclusions

Compared to gold-standard Holter monitoring, the ECG-based Polar H10 chest strap is very accurate at assessing HR over a 24-h time window, while the accuracy of PPG-based Fitbit Inspire 2 tracker is weaker. The results differ between individual patients and between several cardiac patient groups. While neither device is intended or appropriate to replace diagnostic ECG monitoring, both show potential for supporting PA assessment. With current technology, PPG-based HR estimation to assess PA is too inaccurate for use in patients with AF but may be useful in selected other patients with cardiac disease. Our ML-based ARP showed potential to improve PPG-based

HR data by identifying and removing unreliable sections, but its mixed results and the small sample size in this study underscore the need for refinement and independent validation. Further technical improvement may expand the group of patients in whom continuous HR tracking could help assess and guide PA to optimise cardiovascular 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
Katsiaryna Makayed https://orcid.org/0009-0004-2704-9530
Christophe Buyck https://orcid.org/0009-0005-4214-9932
Lieselotte Knaepen https://orcid.org/0000-0003-2816-1896
Juan Sebastian Piedrahita Giraldo https://orcid.org/0000-0002-1691-2915
Sebastiaan Naessens https://orcid.org/0009-0006-9395-2271
Wendy Hens https://orcid.org/0000-0002-9881-6248
Emeline Van Craenenbroeck https://orcid.org/0000-0001-7686-2668
Kris Laukens https://orcid.org/0000-0002-8217-2564
Lien Desteghe https://orcid.org/0000-0001-8641-4658
Hein Heidbuchel https://orcid.org/0000-0001-9301-8127

Ethical considerations

The study was conducted in accordance with ethical guidelines and approved by the Ethics Committee of Antwerp University Hospital (UZA) and the University of Antwerp (UAntwerp) (Central EC reference: 19/21/264, Belgian Unique Number (BUN): B300201941069).

Consent to participate

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

Author contributions

Conceptualisation: all authors; Formal analysis: Paulien Vermunicht; Investigation: Paulien Vermunicht, Lieselotte Knaepen, Sebastiaan Naessens, Wendy Hens; Methodology: all authors; Software: Katsiaryna Makayed, Juan Sebastian Piedrahita Giraldo, Kris Laukens, Paulien Vermunicht; Writing – original draft: Paulien Vermunicht, Katsiaryna Makayed; Writing – review & editing: all authors.

Funding

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 statement

Raw data supporting the conclusions of this article will be made available by the authors upon request.

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: 726.
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. Middeldorp ME, Pathak RK, Meredith M, et al. PREVEntion and regReSsive effect of weight-loss and risk factor modification on atrial fibrillation: the REVERSE-AF study. *Europace* 2018; 20: 1929–1935.
4. Pathak RK, Elliott A, Middeldorp ME, et al. Impact of CARDIOrespiratory FITness on arrhythmia recurrence in obese individuals with atrial fibrillation: the CARDIO-FIT study. *J Am Coll Cardiol* 2015; 66: 985–996.
5. Pathak RK, Middeldorp ME, Meredith M, et al. Long-term effect of goal-directed weight management in an atrial fibrillation cohort: a long-term follow-up study (LEGACY). *J Am Coll Cardiol* 2015; 65: 2159–2169.
6. Wilhelm M. Exercise training and physical activity in patients with heart failure. *Praxis (Bern 1994)* 2018; 107: 951–958.
7. Aune D, Schlesinger S, Leitzmann MF, et al. Physical activity and the risk of heart failure: a systematic review and dose-response meta-analysis of prospective studies. *Eur J Epidemiol* 2021; 36: 367–381.
8. Long L, Mordi IR, Bridges C, et al. Exercise-based cardiac rehabilitation for adults with heart failure. *Cochrane Database Syst Rev* 2019; 1: CD003331.
9. Maurits RD, Bayu FA and Mei CH. Physical activity improves health-related quality of life, 6MWT, and VO2 peak before and during COVID-19 in patients with heart failure: a meta-analysis. *Med Fam-Semergen* 2023; 49: 102039.
10. AbuElkhair A, Boidin M, Buckley BJR, et al. Effects of different exercise types on quality of life for patients with atrial fibrillation: a systematic review and meta-analysis. *J Cardiovasc Med (Hagerstown)* 2023; 24: 87–95. 20221103.
11. Lawler PR, Filion KB and Eisenberg MJ. Efficacy of exercise-based cardiac rehabilitation post-myocardial infarction: a systematic review and meta-analysis of

randomized controlled trials. *Am Heart J* 2011; 162: 571–584.e572.
12. Sartor F, Gelissen J, van Dinther R, et al. Wrist-worn optical and chest strap heart rate comparison in a heterogeneous sample of healthy individuals and in coronary artery disease patients. *BMC Sports Sci Med Rehabil* 2018; 10: 10.
13. 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.
14. Kim C, Song JH and Kim SH. Validation of wearable digital devices for heart rate measurement during exercise test in patients with coronary artery disease. *Ann Rehabil Med* 2023; 47: 261–271.
15. Falter M, Budts W, Goetschalckx K, et al. Accuracy of apple watch measurements for heart rate and energy expenditure in patients with cardiovascular disease: cross-sectional study. *JMIR Mhealth Uhealth* 2019; 7: e11889.
16. Pasadyn SR, Soudan M, Gillinov M, et al. Accuracy of commercially available heart rate monitors in athletes: a prospective study. *Cardiovasc Diagn Ther* 2019; 9: 379–385.
17. 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.
18. 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.
19. Etiwy M, Akhrass Z, Gillinov L, et al. Accuracy of wearable heart rate monitors in cardiac rehabilitation. *Cardiovasc Diagn Ther* 2019; 9: 262–271.
20. Muller AM, Wang NX, Yao J, et al. Heart rate measures from wrist-worn activity trackers in a laboratory and free-living setting: validation study. *JMIR Mhealth Uhealth* 2019; 7: e14120.
21. 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.
22. Zhang A, Xing L, Zou J, et al. Shifting machine learning for healthcare from development to deployment and from models to data. *Nat Biomed Eng* 2022; 6: 1330–1345.
23. Goh CH, Tan LK, Lovell NH, et al. Robust PPG motion artifact detection using a 1-D convolution neural network. *Comput Methods Programs Biomed* 2020; 196: 105596.
24. 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.
25. Athaya T and Choi S. Evaluation of different machine learning models for photoplethysmogram signal artifact detection. In: *2020 international conference on information and communication technology (ICICT)*, 2020, pp.1206–1208. DOI: 10.1109/ICICT50521.2020.00187.
26. 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.
27. Muggeridge DJ, Hickson K, Davies AV, et al. Measurement of heart rate using the polar OH1 and fitbit charge 3 wearable devices in healthy adults during light, moderate, vigorous, and sprint-based exercise: validation study. *JMIR Mhealth Uhealth* 2021; 9: e25313. 20210325.
28. Romagnoli S, Ripanti F, Morettini M, et al. Wearable and portable devices for acquisition of cardiac signals while practicing sport: a scoping review. *Sensors (Basel)* 2023; 23: 20230322.
29. 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.
30. Cardiac Monitors, Heart Rate Meters, and Alarms.
31. Ibrahim NS, Rampal S, Lee WL, et al. Evaluation of wrist-worn photoplethysmography trackers with an electrocardiogram in patients with ischemic heart disease: a validation study. *Cardiovasc Eng Technol* 2024; 15: 12–21. 20231116.
32. Sartor F, Papini G, Cox LGE, et al. Methodological shortcomings of wrist-worn heart rate monitors validations. *J Med Internet Res* 2018; 20: e10108.
33. 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.
34. 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.
35. Al-Kaisey AM, Koshy AN, Ha FJ, et al. Accuracy of wrist-worn heart rate monitors for rate control assessment in atrial fibrillation. *Int J Cardiol* 2020; 300: 161–164. 20191118.
36. Fine J, Branan KL, Rodriguez AJ, et al. Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring. *Biosensors (Basel)* 2021; 11: 20210416.
37. Quinn R, Leader N, Lebovic G, et al. Accuracy of wearable heart rate monitors during exercise in Sinus rhythm and atrial fibrillation. *J Am Coll Cardiol* 2024; 83: 1177–1179.
38. Gorny AW, Liew SJ, Tan CS, et al. Fitbit charge HR wireless heart rate monitor: validation study conducted under free-living conditions. *JMIR Mhealth Uhealth* 2017; 5: e157.
39. Boudreaux BD, Hebert EP, Hollander DB, et al. Validity of wearable activity monitors during cycling and resistance exercise. *Med Sci Sports Exerc* 2018; 50: 624–633. 2017/12/01.
40. Cadmus-Bertram L, Gangnon R, Wirkus EJ, et al. The accuracy of heart rate monitoring by some wrist-worn activity trackers. *Ann Intern Med* 2017; 166: 610–61+.
41. Dooley EE, Golaszewski NM and Bartholomew JB. Estimating accuracy at exercise intensities: a comparative

study of self-monitoring heart rate and physical activity wearable devices. *JMIR Mhealth Uhealth* 2017; 5: 34.

42. 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.
43. Jachymek M, Jachymek MT, Kiedrowicz RM, et al. Wristbands in home-based rehabilitation-validation of heart rate measurement. *Sensors (Basel)* 2021; 22: 20211223.
44. Kroll RR, Boyd JG and Maslove DM. Accuracy of a wrist-worn wearable device for monitoring heart rates in hospital inpatients: a prospective observational study. *J Med Internet Res* 2016; 18: e253.
45. Schwinger RHG. Pathophysiology of heart failure. *Cardiovasc Diagn Ther* 2021; 11: 263–276.
46. Muhammad D, Ahmed I, Ahmad MO, et al. Randomized explainable machine learning models for efficient medical diagnosis. *IEEE J Biomed Health Inform* 2024; 28: 1113.
47. 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.
48. Ventura-Clapier R, Piquereau J, Garnier A, et al. Gender issues in cardiovascular diseases. Focus on energy metabolism. *Biochim Biophys Acta Mol Basis Dis* 2020; 1866: 165722.