Evaluating the accuracy of heart rate sensors based on photoplethysmography for in-the-wild analysis

Liliana Barrios
Department of Computer Science
ETH Zurich
Zurich, Switzerland
liliana.barrios@inf.ethz.ch

Pietro Oldrati
Department of Computer Science
ETH Zurich
Zurich, Switzerland
oldratip@ethz.ch

Silvia Santini
Faculty of Informatics
Università della Svizzera Italiana (USI)
Lugano, Switzerland
silvia.santini@usi.ch

Andreas Lutterotti
Neuroimmunology and Multiple Sclerosis Research
Department of Neurology
University of Zurich
Zurich, Switzerland
andreas.lutterotti@usz.ch

ABSTRACT

Continuous measurement of physiological functions, like heart rate (HR) and heart rate variability (HRV), using commercially available wearable sensors provides the prospects of improving the healthcare of individuals with a positive impact on society, bringing pervasiveness, lower cost, and broader access. However, common wearable devices use photoplethysmography (PPG) to derive data on HR and HRV, and it is yet unclear to which extent PPG signals can be used as a proxy for data collected using medical-grade devices. To address this challenge, we consider five consumer devices to assess the signal quality of HR and two devices measuring HRV and compare them with a standard electrocardiography (ECG) Holter monitor. We collect data from fourteen participants who followed a 55 minutes protocol for at least two sessions. Using this data set, which we make publicly available to the research community, we show that PPG is a valid proxy for both HR and standard time- and frequency-domain measurements of HRV. Further, we demonstrate that wearable devices are suitable for monitoring both HR and HRV in daily life but might be limited during strenuous exercise. The study indicates that armband-based devices are more reliable than wrist-based wearables for HRV assessment.

CCS CONCEPTS

• **Hardware → Sensor applications and deployments**; **Sensor devices and platforms**;

KEYWORDS

Mobile health, Wearable, Heart rate, Heart rate variability, Validation, Photoplethysmography

PervasiveHealth'19, May 20–23, 2019, Trento, Italy

ACM ISBN 978-1-4503-6126-2/19/05…$15.00
https://doi.org/10.1145/3329189.3329215

ACM Reference Format:
Liliana Barrios, Pietro Oldrati, Silvia Santini, and Andreas Lutterotti. 2019. Evaluating the accuracy of heart rate sensors based on photoplethysmography for in-the-wild analysis. In *The 13th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth'19), May 20–23, 2019, Trento, Italy*. ACM, Trento, Italy, 11 pages. https://doi.org/10.1145/3329189.3329215

1 INTRODUCTION

Market shipment of wearable devices is expected to reach 190.4 million units by 2022, according to the latest report of the International Data Corporation (IDC), representing an increase of 60.6% from the 115.4 million units shipped in 2017 [20]. Rapid advances in sensor technology contribute to the increasing interest in wearables. They are no longer perceived solely as fitness trackers able to count steps, but also as an opportunity to enhance health-care systems [21]. Health care can significantly benefit from wearable technology by allowing continuous monitoring [4, 26, 27], as opposed to the traditional, often limiting, one-time assessments that evaluate the health status of patients during a medical consultation. Moreover, wearables are unobtrusive, cheaper than standard medical devices, and can measure body vital signs such as heart rate (HR) and derive further parameters such as heart rate variability (HRV). Even though wearable devices offer several advantages, their level of agreement with conventional medical-grade devices is still questionable. A significant limitation of existing validation studies is that they rely solely on linear correlation, which has been shown to be insufficient to evaluate agreement [6, 14, 35], or fail to report the intraclass correlation coefficient (ICC) form used in their analysis [25]. In this paper, we make use of the Bland-Altman [6] (BA) plot and report ICC with its corresponding model, type, and definition. Both BA and ICC are widely considered as appropriate methods for validating agreement between two devices [6, 14, 25, 35].

HRV is a powerful metric for evaluating the health status of a person. In particular, it is widely used for assessing the role of the autonomous nervous system in healthy individuals and patients [1]. It is a good prognostic tool for myocardial infarction [7], atrial fibrillation [28], and risk of mortality [37]. Moreover, it can help to assess conditions like stress [9], depression [23], and fatigue [33].

Routine heart rate evaluation is performed through electrocardiography (ECG). ECG records the electrical activity of the heart over a period of time using electrodes placed over the skin [24]. Intervals between successive heartbeats are known as inter-beat (RR) Intervals. HRV metrics are derived from those intervals. An ECG Holter monitor is commonly used to evaluate RR-intervals in ambulatory conditions [1, 35]. This device typically allows the monitoring of the heart rate at high resolution, i.e. 250-1000 Hz, over a period of 24-72 hours.

In the last decades, standalone heart rate monitors based on photoplethysmography (PPG) gained popularity. These devices have the advantage that they do not require additional equipment such as chest straps. Today, it is common for consumer-wearable devices such as smartwatches and fitness trackers to include a PPG heart rate monitor.

PPG is a non-invasive technique for monitoring blood volume changes in the microvascular bed of tissue [8]. PPG technology detects the blood pulse wave by illuminating the skin and measuring the portion of light which is reflected back to the device. Heart rate is computed by detecting peaks (beats) in the PPG signal. A parameter that can be extracted from the PPG data is the heart rate variability (HRV). HRV measures the variation in the time interval between heartbeats.

Wearables capable of measuring RR-intervals provide many advantages over traditional methods, e.g., they are portable, easy to use, low-cost and do not require specialized knowledge for the placement of the electrodes. A number of studies [22, 38] analyze the level of agreement between the mean HR of these devices and ECG Holters. However, the level of agreement of the RR-intervals and HRV metrics has rarely been subjected to study. Different studies show that the chest strap monitors used by Polar devices have good agreement with the RR-intervals derived from ECG Holters with 3, 5, and 12 leads [3, 16, 18, 35].

The European Society of Cardiology and the North American Society of Pacing and Electrophysiology highlight that it is essential to validate the potential of new technologies to accurately and reliably record RR-intervals for use in clinical or research applications [1]. Moreover, a recent study on the ability of wearable devices to measure HRV concludes that there is a need for robust studies in non-stationary conditions with appropriate methodologies to assess the accuracy of HRV derived from consumer-wearable devices with PPG heart rate sensors [14]. Thus, part of our study aims at verifying if the RR-intervals obtained from PPG sensors in consumer-wearable devices are a valid proxy for HRV in non-stationary conditions. To this end, we conduct a series of experiments to evaluate the level of agreement between consumer-wearable devices capable of measuring heart rate and RR-intervals based on PPG technology. The inclusion criteria for sensors in our validation study are: availability, access to raw data, capable of measuring HR, inter-beat interval (IBI) and electrodermal activity (EDA), battery life for over 24 hours and storage capability for over a day. These considerations would enable monitoring over extended periods. Following our criteria we select two devices, the Empatica E4 [10] and Everion [5]. As the baseline, we consider a medical-grade 5-lead ECG Holter [12]. Furthermore, we explore how the level of agreement of these two devices compares with the mean heart rate derived from popular fitness trackers: Fitbit Charge HR [11], Polar OH1 [34], and Wahoo Ticker Fit [40][1]. Finally, we perform a quantitative analysis of participants placement preferences of the wearable device during long-term monitoring.

In summary, this paper makes the following contributions:

- Evaluation of level of agreement of five different PPG heart rate monitors (Empatica E4, Everion, Fitbit Charge HR, Polar OH1, Wahoo Ticker Fit) and a standard Holter monitor (Seer 1000).
- Evaluation of level of agreement of HR, RR-Intervals and HRV metrics derived from consumer-wearable PPG monitors (Empatica, Biovotion) and a standard Holter monitor (Seer 1000).
- Activity dataset[2] comprised of fourteen participants (7 female and 7 male), six different wearable devices including HR, IBI, EDA, and three different experiments (30 sessions of 55 minutes each).

The goal of this study is to provide an analysis of the accuracy of new consumer-wearable devices. To the best of our knowledge, the resulting dataset, which we make publicly available through the ETH research collection, is the first of its kind. Our contributions are thus especially relevant for researchers interested in exploring the possibilities of using the HR and EDA sensors included on wearable devices.

2 BACKGROUND AND RELATED WORK

The Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology [1] released a report in an effort to provide standardization in the research and application of HRV [1]. HRV metrics are extracted from the ECG signals. The largest-amplitude portion of the ECG signal is called QRS complex [29] and corresponds to the depolarization of the right and left ventricles of the heart. Normal-to-normal (NN) intervals, are the intervals between adjacent QRS complexes. HRV refers to the oscillation in the interval between consecutive heartbeats (RR intervals) and oscillation between instantaneous heart rate.

Validation studies on HR and HRV derived from wearable devices. Several studies show that wearable devices can accurately measure mean HR based on PPG [22, 32, 41]. Others focus on assessing the accuracy of HRV metrics extracted from chest strap monitors. Many authors [3, 16, 18, 31] have used the Polar chest strap, which is an electrode-based sensor, in their studies. For instance, Giles et al. [16] show that the Polar V800 is able to produce RR interval recordings consistent with an ECG during rest, and that the HRV parameters derived from these recordings are also highly comparable. Nunan et al. [31] compare the number of RR intervals recorded by the Polar S810 and a standard 12-lead ECG monitor and found that both devices have good agreement when the wearer is laying down. Additionally, they found good agreement between the derived HRV metrics. Hernando et al. [18] explore the reliability of Polar RS800 to measure HRV metrics during exercise. Their work shows that at high exercise intensity low-frequency domain measurements have excellent reliability indices, however, high-frequency measurements have low agreement.

[1]Disclaimer: the authors have no conflict of interest nor received funding from any of these device manufacturers

[2]https://www.research-collection.ethz.ch/

Table 1: Collected sensor data.

Device	Variable	Frequency	Export
Empatica E4	HR (bpm)	1 Hz	Empatica
	IBI	-	
Everion	HR (bpm)	1 Hz	Bluetooth LE
	IBI	-	
Seer 1000	HR (bpm)	1 Hz	CardioDay V2.5
	RR-intervals	-	
Polar	HR (bpm)	1 Hz	PolarFlow
Fitbit	HR (bpm)	1/3 Hz (varies)	Fitbit.com
Wahoo	HR (bpm)	1 Hz	Wahoo Fitness

Less common are studies that examine the validity of HRV metrics derived from off-the-shelf wearable devices with PPG technology. Giardino et al. [15] found good agreement between the HRV metrics obtained from a finger plethysmograph and an ECG with three leads. Vescio et al. [39] developed a customized device that converts the PPG signal generated by a LED-photodiode couple placed on the earlobe into electric pulses. Their device was tested under stationary conditions with 10 participants. Their results show good agreement with the ECG recordings. On a recent review about heart rate variability based on wearable device, Georgiou et al. [14] reviewed 308 articles, from those only two articles considered measuring HRV with wearable devices using PPG technology. Their research concludes that there is a need for more robust studies in non-stationary conditions, with appropriate methodology, acquisition and analysis techniques to evaluate the ability of wearables to measure HRV based on PPG [14].

In summary, previous research has shown that the interbeat intervals (IBI) derived from PPG signals are comparable to the RR intervals obtained from ECG Holter monitors under non-ambulatory conditions. However, little is known about the quality of the signal provided by off-the-shelf wearable device and their capability to measure HRV [14]. Therefore, in this work we address this open question by evaluating two devices (Everion and Empatica) under different conditions.

3 METHODS

Our goal is to evaluate the performance of PPG sensors, found in commodity wearable devices, under different settings for measuring HR and IBI. To this end, we conducted a series of experiments. Fourteen volunteers, seven males with median age 33 (range 23-54) and seven females with median age 36 (range 26-51), took part in the study. Their mean height is 170 cm and mean weight is 67 kg. Volunteers gave full written informed consent to participate in the study. All procedures were approved by the ETH Zurich local committee (EK 2018-N-89).

Table 1 shows details regarding the used data, frequency, and export method. During the experiment, Polar data was stored locally on the devices and later exported using the Polar Flow smartphone application. Similarly, Wahoo data was exported to a CSV file using the Wahoo Fitness application. Fitbit data was downloaded from fitbit.com [17]. Empatica data was stored locally on the devices and later exported with the Empatica Connect software. Everion data was streamed via Bluethooth Low Energy to an Android phone during the experiment and later exported as a CSV file. Finally, QRS complexes were obtained from the ECG Holter with the software CardioDay [13] from GE Healthcare.

In order to assess the validity of the PPG sensors, we use a protocol similar to Jo et al. [22]. The protocol starts with 5 minutes of resting on a stationary bike followed by five activities: biking (60 W), biking (120 W), walking (5 km/h), jogging (8 km/h) and running (10 km/h). Each activity lasts for 5 minutes, and between each activity, there is a resting period of 5 minutes. The left side of Figure 1 depicts our validation protocol.

Experiment I - Accuracy of PPG based HR monitors. The goal of this experiment is to compare the level of agreement of the Empatica E4 [10](version 1) and Everion [5](VSM1-3.0, M4 version 03.11.00) with the mean HR derived from popular fitness trackers: Fitbit Charge HR [11], Polar OH1 [34], and Wahoo Ticker Fit [40]. Participants wore two Empatica E4 devices (one on each wrist), two Everion devices (one per arm), and a medical-grade Holter monitor, the General Electric Seer 1000 [12] with 5 leads, as depicted on the right side of Figure 1. The fitness trackers were placed on the arm of the participants without a predefined position. Six participants took part in this experiment, three male and three female. Each participant completed our validation protocol two times on different days.

Experiment II - Comparing Everion, Empatica and Holter . The goal of this experiment is to assess the heart rate and interbeat intervals derived from PPG sensors as a valid proxy for HRV. We considered two off-the-shelf sensors capable of measuring HR through photoplethysmography (PPG) and electrodermal activity (EDA): Empatica E4 and Everion devices. Fourteen participants took part in the experiment and completed the protocol two times on different days. Participants wore two Empatica E4 devices (one on each wrist), two Everion devices (one per arm), and a medical-grade Holter monitor, the General Electric Seer 1000 [12] with 5 leads, to record ECG signals. The sensor placement is depicted on the right side of Figure 1. Moreover, we explore the variance between successive measurements. To this end, two participants performed three extra sessions. We compute the mean HR difference between Everion, Empatica and Holter per activity and perform an ANOVA analysis per session.

4 DATA ANALYSIS

Before evaluating the level of agreement of the different devices involved in our experiments, sequences of interbeat intervals derived from the ECG and PPG devices were aligned through cross correlation mechanism. Additionally, we did an outlier analysis and excluded data points resulting from potential errors or artifacts caused during data acquisition, i.e. HR equal to zero during the experiment.

4.1 Metrics

We use different metrics to measure the performance and level of agreement of the different devices. We report mean and standard deviation of the HR. We evaluate the existence of bias, with its limits of agreement [LoA], using the Bland-Altman [6] plot. The Bland-Altman plot [6] is a plot of the difference between two methods

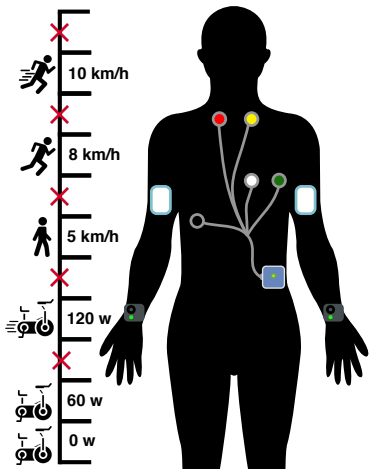

Figure 1: Sensor validation protocol at the left side and sensor placement at the right side. Empatica E4 devices placed on the subject's wrist; Everion devices on the arms; and the Holter monitor attached with five electrodes to the chest.

against their mean, allowing to investigate any possible relationship between the measurement error and the true value. In this plot none of the values are considered to be the true value, thus, the mean value is used as the best estimate. In our analysis, we consider the HR derived from the Holter versus HR derived from the wearable devices. Additionally, we compute the intraclass correlation coefficient (*ICC*) with its 95% confidence interval, Pearson correlation (*corr*) and squared error R^2.

Following Koo and Li guidelines for selecting and reporting ICC [25], we computed ICC and its 95% confident intervals using IBM SPSS statistics [2] based on single measurement type, absolute agreement definition and 2-way mixed-effects model. ICC results are interpreted as in [25]: values less than 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability.

4.2 Heart rate variability

We derive different time and frequency domain measures of HRV from the IBI and NN time series provided by the Empatica, Everion and ECG Holter. An overview of the metrics is depicted in Table 2. According to the recommendations [1], 5 minutes is an appropriate length for short-term recordings of HRV. When analyzing the spectrum for short-term recordings time varies between 2 and 5 minutes. We use fast Fourier transformation (FFT) to derive frequency domain HRV measurements from the IBI interval time series. In accordance with [1], we divide the power spectrum for frequency domain HRV analysis into the following bands: VLF (0.00 - 0.04 Hz), LF (0.04 - 0.15 Hz) and HF (0.15 - 0.40 Hz). For the calculation of HRV parameters, we select identical segments larger than 180 s of NN intervals from the ECG and a wearable device. Then, we apply cubic interpolation. Finally, we analyze the spectrum with Welch's periodogram using the following parameters: hamming window, overlap of 50% and linear detrend.

Table 2: Overview of the heart rate variability metrics computed.

Metric	Domain	Definition
RMSS	time	Square root of the mean squared differences of successive NN intervals.
SDNN	time	Standard deviation of the NN intervals.
NN50	time	Number of interval differences of successive NN intervals greater than 50 ms.
pNN50	time	Proportion derived by dividing NN50 by the total number of NN intervals.
VLF	frequency	Very low frequency.
LF	frequency	Low frequency.
HF	frequency	High frequency.
LFnu	frequency	Normalized low frequency.
HFnu	frequency	Normalized high frequency.
LF:HF	frequency	Ratio.

5 RESULTS

In this section, we start by showing the accuracy of the HR sensors in different off-the-shelf wearable devices. Following, we look into their ability to extract IBI.

5.1 Accuracy of PPG based HR monitors

We compare the performance of the Empatica and the Everion versus commonly used fitness trackers. Table 3 depicts different metrics comparing each wearable with the Holter. The table is organized by device with its associated sample size. Additionally, we split each case into activities. To make a fair comparison between the fitness trackers, Empatica and Everion, we did not use quality filters on the data as this functionality is not available on the fitness trackers. Figure 2 depicts the Bland Altman plot for all devices.

5.1.1 Wrist-based devices. Figure 2 shows that the wrist-based devices, Empatica and Fitbit, have the largest bias, 17.35 bpm and 5.898 bpm respectively. The analysis per activity (Table 3) shows that in both cases the bias increases with the level of activity. The Fitbit is more affected during bike activities and the Empatica during jogging and running. Figure 3 shows an overview of the level of agreement of each device per activity. These results are consistent with the bias analysis. Wrist-based devices are more affected than armband-based devices. The Empatica shows lowest agreement (poor agreement) within all different activities, especially during jog and run. However, the results for Empatica are improved after filtering the data, resulting in good agreement during the initial, and rest activities. The Fitbit has the lowest agreement during bike activities, good agreement during rest and moderate agreement during the initial activity. A possible explanation for the poor agreement during bike activities with these devices is the posture of the wrist on the bike. Bending of the wrist can generate loose contact between the skin and the heart rate monitor resulting in low quality measurements.

5.1.2 Armband-based devices. In Figure 2 we can observe that the bias and data distribution is similar for all armband-based devices,

Table 3: Experiment I - Heart rate analysis per activity

Device	Activity	Size	Mean/Seer	STD/Seer	ICC [95% CI]	Corr	Bias [95% LoA]
Empatica	init	6268	84.21/87.04	9.91/13.60	.464 [+.422,+.502]	0.50	+2.83 [-21.04, +26.69]
(79241)	rest	33676	92.94/100.15	15.87/18.61	.466 [+.349,+.558]	0.51	+7.21 [-26.46, +40.88]
	bike (60 W)	7856	90.66/108.61	21.23/12.88	.167 [-.002,+.314]	0.29	+17.95 [-24.07, +59.96]
	bike (120 W)	7968	104.54/137.18	36.99/20.56	.223 [-.027,+.422]	0.42	+32.64 [-33.95, +99.24]
	walk	8096	102.31/108.91	16.09/16.50	.433 [+.331,+.517]	0.47	+6.60 [-26.32, +39.52]
	jog	7690	102.98/138.73	26.28/18.29	.026 [-.015,+.068]	0.06	+35.76 [-25.11, +96.63]
	run	7687	102.85/152.94	30.08/19.74	.016 [-.014,+.046]	0.05	+50.08 [-18.77 ,+118.94]
	avg		**97.21/119.08**	**22.35/17.17**	**0.256 [+.149, +.346]**	**0.33**	**+21.87 [-25.10, +68.84]**
Everion	init	6268	87.71/87.04	14.26/13.60	.957 [+.953,+.960]	0.96	-0.67 [-8.59, +7.25]
(78821)	rest	33590	101.19/100.13	19.36/18.62	.972 [+.967,+.976]	0.97	-1.06 [-9.59, +7.47]
	bike (60 W)	7856	108.88/108.61	12.89/12.88	.981 [+.980,+.982]	0.98	-0.27 [-5.15, +4.61]
	bike (120 W)	7666	137.58/137.40	20.14/20.55	.988 [+.988,+.989]	0.99	-0.18 [-6.31, +5.95]
	walk	8096	109.14/108.91	16.61/16.50	.993 [+.993,+.993]	0.99	-0.23 [-4.05, +3.58]
	jog	7683	137.87/138.75	17.98/18.28	.965 [+.962,+.969]	0.97	+0.89 [-8.29, +10.07]
	run	7662	152.66/152.94	19.84/19.77	.952 [+.950,+.954]	0.95	+0.29 [-11.73, +12.30]
	avg		**119.29/119.11**	**17.30/17.17**	**0.972 [+.970,+.974]**	**0.97**	**-0.17 [-7.67, +7.32]**
Fitbit	init	3134	83.23/87.04	11.48/13.60	.767 [+.645,+.838]	0.82	+3.80 [-11.34, +18.95]
(39624)	rest	16812	98.71/100.14	19.81/18.62	.827 [+.821,+.832]	0.84	+1.43 [-20.21, +23.07]
	bike (60 W)	3928	99.02/108.61	16.56/12.88	.499 [+.170,+.682]	0.63	+9.59 [-16.21,+35.38]
	bike (120 W)	3984	118.37/137.18	27.97/20.56	.353 [+.079,+.542]	0.48	+18.81[-31.24, +68.87]
	walk	4048	103.48/108.90	12.73/16.50	.729 [+.541,+.824]	0.81	+5.43 [-13.62, +24.47]
	jog	3854	130.61/138.70	20.01/18.29	.703 [+.441,+.822]	0.78	+8.09 [-17.09, +33.27]
	run	3852	144.74/152.96	22.01/19.73	.782 [+.468,+.887]	0.86	+8.22 [-13.93, +30.38]
	avg		**111.17/119.08**	**18.63/17.17**	**0.665 [+.452, +.776]**	**0.74**	**+7.91 [-17.66, +33.48]**
Polar	init	3134	88.12/87.04	14.44/13.60	.959 [+.936,+.953]	0.95	-1.08 [-9.93, +7.77]
(39624)	rest	16812	101.60/100.14	19.45/18.62	.969 [+.959,+.976]	0.97	-1.46 [-10.27, +7.35]
	bike (60 W)	3928	108.73/108.61	13.30/12.88	.972 [+.970,+.973]	0.97	-0.13 [-6.24,+5.99]
	bike (120 W)	3984	136.90/137.18	21.23/20.56	.984 [+.983,+.985]	0.98	+0.28 [-7.10, +7.66]
	walk	4048	108.89/108.90	16.74/16.50	.989 [+.988,+.989]	0.99	+0.01 [-4.86, +4.89]
	jog	3854	137.51/138.70	19.17/18.29	.964 [+.957,+.969]	0.97	+1.19 [-8.41, +10.80]
	run	3852	152.27/152.96	21.09/19.73	.950 [+.947,+.953]	0.95	+0.69 [-11.89,+13.26]
	avg		**119.15/119.08**	**17.92/17.17**	**0.969 [+.963,+.971]**	**0.97**	**-0.07 [-8.38, +8.24]**
Wahoo	init	3094	84.91/86.97	13.77/13.51	.913 [+.875,+.936]	0.92	+2.06 [-8.41,+12.53]
(38492)	rest	16406	99.37/99.95	19.52/18.76	.965 [+.964,+.967]	0.97	+0.58 [-9.22,+10.38]
	bike (60 W)	3567	106.16/106.76	11.80/11.81	.971 [+.966,+.974]	0.97	+0.60 [-4.88, +6.08]
	bike (120 W)	3732	135.42/136.22	20.87/20.64	.984 [+.982,+.986]	0.99	+0.80 [-6.21, +7.81]
	walk	3975	107.80/108.94	17.00/16.63	.972 [+.964,+.977]	0.97	+1.14 [-6.36 ,+8.65]
	jog	3854	136.58/138.70	19.67/18.29	.939 [+.918,+.952]	0.95	+2.13 [-10.29, +14.54]
	run	3852	151.65/152.96	20.88/19.73	.937 [+.930,+.944]	0.94	+1.30 [-12.56, +15.18]
	avg		**117.41/118.64**	**17.64/17.05**	**0.954 [+.943,+.962]**	**0.96**	**+1.23 [-8.28, +10.74]**

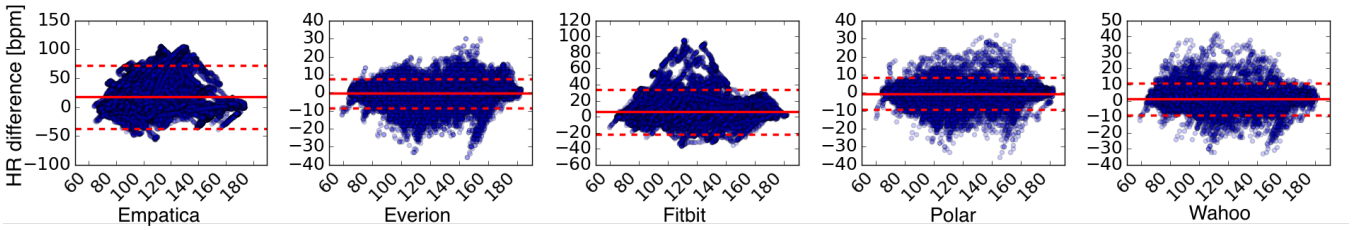

Figure 2: Bland-Altman plot with LoA for each HR monitor. Wrist-based devices show largest bias (Empatica 17.35 [-37.16, +71.86] and Fitbit 5.89 [-22.19, +33.97]) than armband-based monitors (Everion -0.46 [-8.67, +7.75], Polar -0.51 [-9.38, 8.36], and Wahoo 1.01 [-8.95, +10.96]).

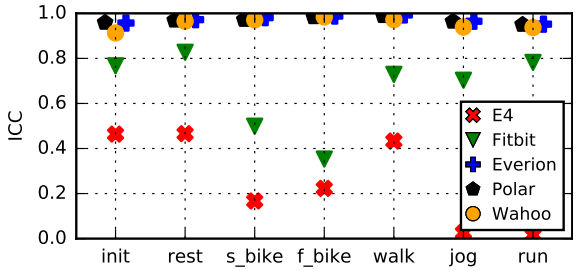

Figure 3: Level of agreement according to ICC for each device in experiment II. Notably the level of agreement of the wrist-based devices is lower than for the armband-based devices with Empatica being more affected as the activity level increases.

i.e., Everion, Polar, and Wahoo. The three devices show smaller bias in comparison to the wrist-based devices. There is no particular trend on the bias depending on the activity. Similarly, Figure 3 shows that all devices have a similar level of agreement in terms of ICC. Everion, Polar and Wahoo show excellent reliability with regards to the Holter in all the activities, showing that armband-devices are less susceptible to artifacts due to movement.

5.1.3 Users' preferences. After the experiments, participants completed a short questionnaire indicating their preferred style of wearable device (armband, wristband) for continuous monitoring during (i) day, (ii) sleep and (iii) 24/7. A Cochran's Q test did not indicate any differences among the three proportions, p = .717, showing that user's preference is not affected by the duration of the monitoring phase.

5.1.4 Discussion. In general, devices perform better when there is less movement involved, as in the case of the rest period and initial activity. Both wrist-located devices perform poorly on the bike, jog and run activities, indicating that (i) wrist's posture may affect the accuracy of wrist-based monitoring, (ii) wrist-based monitors are more susceptible to movement in comparison with non-wrist devices such as the Polar OH1, Wahoo Ticker, and Everion. From our short questionnaire, we learned that 58.3% of the participants prefer wrist-based devices. Thus, when designing experiments, there is a trade-off to be made between comfort/users' preferences and reliability.

5.2 Comparing Everion, Empatica and Holter

To compare both devices, we started by computing metrics corresponding to the mean HR derived from the Empatica E4 and Everion relative to the medical-grade Holter. Both devices, Empatica and Everion, have mechanisms to assess the quality of the retrieved heart rate i.e., low or high quality. We refer to high quality datasets as *Everion Best* and *Empatica Best*. To filter values depending on their quality, Everion provides a heart rate quality parameter. In our analysis *Everion Best* corresponds to HR quality parameter of 99%. Additionally, we include in our analysis a dataset with HR quality 90%, we refer to these as *Everion q90*. In the case of Empatica, we consider HR quality to be high when IBI is present on the data. Empatica applies a filter to its IBI data, thus wrong beats are

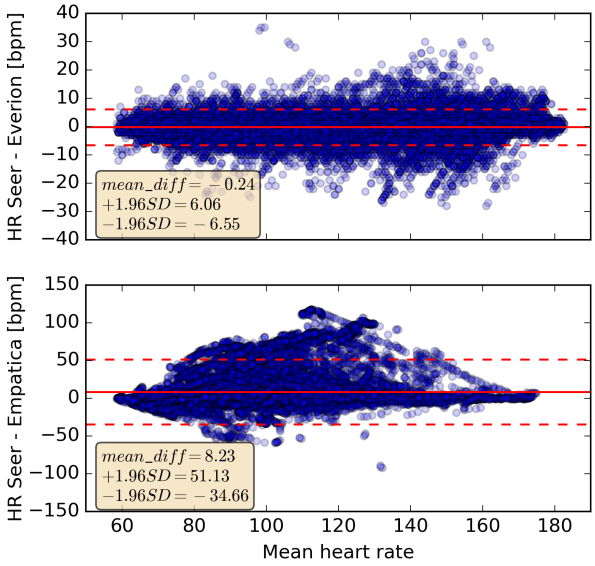

Figure 4: Bland-Altman plot for Empatica/Holter and Everion/Holter. Empatica's mean bias (8.23 LoA [51.13,-34.66]) is larger than Everion's mean bias (-0.24 LoA [6.06, -6.55]). Overall the data seems to be well distributed showing no particular pattern.

not included in the output IBI file [19]. Table 4 shows the metrics corresponding to the datasets *Everion Best* and *Empatica Best*.

Figure 4 shows the Bland-Altman plot for the *Empatica Best* and *Everion Best* datasets. For Empatica, the 95% limits of agreement ranged from -34.66 to +51.13 with a mean difference of 8.23 bpm. While for the Everion, the 95% limits of agreement ranged from -6.55 to +6.06 with a mean difference of -0.24 bpm. The data distribution of both devices shows no specific pattern. Further analysis showed that the data of both devices is normally distributed with the majority of the data points located within two standard deviations.

5.2.1 Performance per activity. To understand why the Empatica has a larger bias, Figure 5 shows the bias per activity. The bias increases significantly during activities involving exercise (μ bias 30 bpm) and remains low (μ bias 1 bpm) during the initial activity, rest and walking. Similarly, during non-strenuous activities the Empatica's mean HR and standard deviation are similar to the Holter's. However, as the level of activity increases the difference between the Empatica's mean HR and the Holter also increases. This behavior is not observed with the Everion. Figure 5 shows that over all activities the Everion's mean HR behaves similarly to the Holter, with no significant difference between the values. Additionally, we can observe that the Everion's bias increases as the level of activity increases. The largest bias occurs during rest, where Everion underestimate the mean HR in average by 1 bpm. However, over all activities the bias of the Everion is small with a mean value of -0.01 bpm.

Figure 6 shows at the bottom the ICC per activity for *Everion Best*, *Everion q90* and *Empatica Best*. The Empatica's ICC significantly

Table 4: Experiment II - Heart rate analysis per activity

Case	Activity	Size	Mean/Seer	STD/Seer	ICC [95% CI]	Corr	Bias [95% LoA]
Everion Best (63592)	init	3808	83.09/82.83	14.03/13.94	.979 [+.978,+.981]	0.98	-0.26 [-5.83, +5.31]
	rest	21364	103.28/102.21	22.91/22.67	.988 [+.984,+.991]	0.99	-1.07 [-7.67,+5.53]
	bike (60 W)	9445	106.87/106.92	13.84/13.73	.989 [+.989,+.990]	0.99	+0.05 [-3.91, +4.01]
	bike (120 W)	8299	137.37/137.54	21.72/21.72	.995 [+.995,+.995]	1.00	+0.17 [-4.00, +4.34]
	walk	6722	104.00/103.90	16.57/16.54	.995 [+.995,+.995]	0.99	-0.10 [-3.36, +3.16]
	jog	8093	137.32/137.66	17.24/17.22	.973 [+.971,+.974]	0.97	+0.34 [-7.55, +8.23]
	run	5861	151.27/152.03	21.41/20.76	.974 [+.971,+.976]	0.97	+0.77 [-8.59, +10.12]
	avg		**117.60/117.59**	**18.25/18.08**	**0.985 [+.983,+.986]**	**0.99**	**-0.01 [-5.84,+5.81]**
Empatica Best (35705)	init	3852	81.63/81.64	11.56/12.83	.755 [+.741,+.769]	0.76	+0.01 [-16.74, +16.76]
	rest	20883	94.65/96.01	16.44/18.22	.834 [+.825,+.842]	0.84	+1.36 [-18.09, +20.82]
	bike (60 W)	4948	88.17/108.33	20.33/15.43	.031 [-.006,+.067]	0.05	+20.16 [-28.61, +68.94]
	bike (120 W)	4401	103.46/137.93	34.33/22.11	.118 [-.023,+.247]	0.22	+34.47 [-37.06, +106.00]
	walk	1287	104.02/106.43	16.46/18.03	.597 [+.558,+.634]	0.61	+2.41 [-27.74, +32.56]
	jog	233	107.95/140.09	22.33/22.17	.073 [-.044,+.195]	0.15	+32.14 [-24.77, +89.04]
	run	101	112.02/146.14	31.32/ 22.87	.120 [-.056,+.300]	0.22	+34.12 [-33.35, +101.59]
	avg		**98.84/116.65**	**21.83/18.81**	**0.361 [+.285,+.436]**	**0.41**	**17.81 [-26.62,+62.24]**

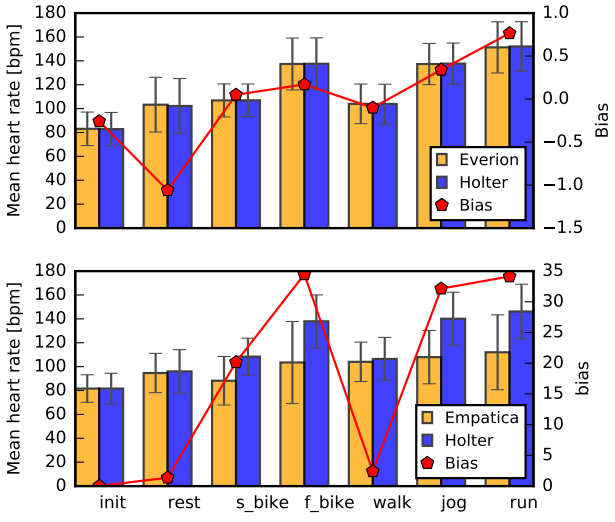

Figure 5: Mean HR, standard deviation and bias per activity of Empatica, Everion and Holter. In particular, Empatica's bias increases significantly during strenuous activities and remains low while being less active. The difference between Empatica's mean HR and Holter shows a similar behavior to the bias, increasing with exercise. Everion's bias increases with the level of activity but overall remains low. Everion's mean HR and standard deviation show similar behaviors as the Holter.

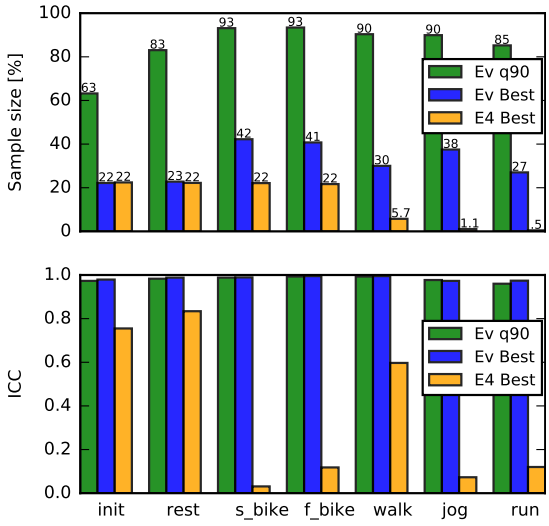

Figure 6: The bottom image shows the ICC corresponding to each activity. In particular, Empatica's accuracy is significantly lower during strenuous activities. In the case of Everion, both datasets are comparable showing high ICC over all activities. The top figure shows the fraction of the sample size of each dataset in relation to its original dataset. The Everion q90 dataset is up to four times the size of Everion Best. Empatica best is considerably smaller than its original dataset, and it is more affected while jogging and running.

decreases during the bike, jog and run activities, indicating that the device is not suitable for monitoring HR during strenuous activity. The Everion, on the other hand, shows high ICC in both datasets. Moreover, the top of Figure 6 shows that the *Everion q90* dataset is around three times larger than *Everion Best*. Thus, relaxing the heart rate quality threshold allows to have a larger dataset with similar accuracy. In the case of Empatica, we can see that the size of the dataset gets greatly affected as movement increases with the sample size being only 0.5% of the original data for the run activity.

Finally, we analyzed the mean HR difference per activity on five successive sessions and found no statistically significant differences between group means as determined by one-way ANOVA F(4,30) = .338, p = .850. In summary, both devices provide useful mechanisms

to ensure a high quality of the resulting dataset. We recommend to make use of these parameters. Overall we consider both of the Everion's datasets comparable and showing excellent agreement with respect to the Holter in all activities. Empatica provides good agreement for the initial activity and rest periods, and moderate agreement while walking. Our results indicate that Empatica is less suitable for tracking mean HR during ambulatory conditions or high intensity activities.

5.3 Heart rate variability analysis

We started our analysis by extracting IBI segments from the time-series. From the Empatica we obtained a total of 137 IBI segments with an average length of 49 s. Per activity we gathered the following number of segments (with mean duration): init 12 (μ 44 s), rest 81 (μ 50 s), bike slow 23 (μ 44 s), bike fast 20 (μ 53 s), walk 1 (μ 34 s). Only two IBI segments in the whole dataset are longer than 2 minutes. The recommended length for short-term HRV analysis ranges from 3-5 minutes [1, 36]. Thus, we are unable to compute short-term HRV analysis for this device. Future work can overcome this limitation by applying techniques to approximate the missing IBI signal.

Table 5 depicts the results of our HRV analysis comparing the Everion with the ECG Holter. The table is organized per activity. For each activity we extracted IBI segments larger than 3 minutes. Figure 7 depicts the level of agreement between the HRV metrics derived from the Everion and Holter during each activity. There is good agreement during the initial and rest activities in all metrics. Agreement decreases with higher level of activity and it varies depending on the metric. HF is more affected with increasing level of activity.

5.3.1 Sedentary activities. During the initial and rest activity there is good agreement in all HRV measurements. For the initial activity, we found 17 segments larger than 3 minutes, the mean length of the segments is 241 s. Highest agreement occurs on the frequency domain metric LF with *ICC* between +.952 and +.993, indicating excellent agreement. Lowest agreement occurs on the ratio LF:HF with mean *ICC* ranging from +.215 to +.846, indicating poor agreement. Time domain measurements indicate better agreement with the Holter, ranging from moderate to excellent. For the rest activity, we collected 111 segments with an average duration of 249 s. Overall the results are satisfactory in this activity. Excellent agreement occurs in all time domain metrics and LF. Followed by good agreement in HF, moderate agreement in the normalized LF and HF, and poor agreement for the ration LF:HF.

5.3.2 Moderate/High intensity activities. In the activity bike (60 W) the average length of the 28 considered segments is 294 s. Highest agreement happens in LF with *ICC* ranging from +.886 to +.974, indicating good agreement. Followed by moderate agreement in SDNN with *ICC* ranging from +.596 to +.899. The rest of the metrics show poor agreement in relation to the Holter. In the bike (120 W) activity only 13 segments are larger than 3 minutes, with an average length of 272 s. In this activity only one metric shows moderate agreement, SDNN with *ICC* ranging from +.552 to +.946. The rest of the metrics show poor agreement. We consider 17 segments with average length of 291 s for the walk activity. In this activity all metrics show poor reliability. The activities jog and run are not included in the analysis as we were unable to extract IBI segments longer than 3 minutes from them.

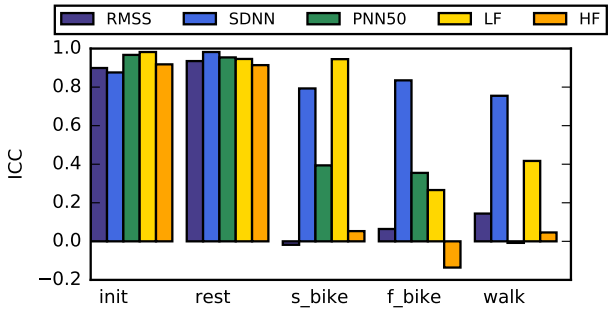

Figure 7: Intraclass correlation for different HRV metrics. As the activity level increases the ICC decreases, more notably in the HF band.

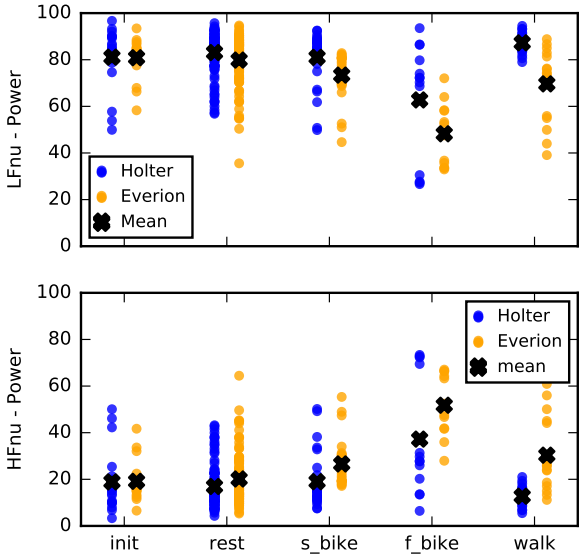

Figure 8: LFnu increases during low-moderate intensity exercise and decreases during higher intensity exercise, while HFnu demonstrates the opposite behavior.

5.3.3 Discussion. The evaluation of HRV during ambulatory activities is still subject of study. LF shows higher agreement than HF as the activity level increases. This is consistent with the results found in [18]. Moreover, our findings agree to those reported in [30]. In general, the reliability of time domain and frequency domain measurements decrease as exercises intensity increases. Moreover, our analysis shows similar results in terms of the normalized spectral analysis, with LFnu increasing during low-moderate intensity exercise and decreasing during higher intensity exercise, while HFnu shows the opposite response as shown in Figure 8. Additionally, Figure 8 shows that the normalized metrics of the Everion and the

Table 5: Experiment II - Heart rate variability analysis per activity for the Everion device.

Activity	Metric	Mean/Seer	STD/Seer	ICC [95% CI]	Corr	R^2	Bias [95% LoA]
Init	RMSS	23.17/23.39	12.40/14.92	+.899 [+.742, +.967]	+0.91	+0.82	+0.22 [-12.22, +12.65]
μ len: 241 s	SDNN	58.49/60.62	16.71/17.54	+.876 [+.697, +.953]	+0.88	+0.75	+2.13 [-14.57, +18.84]
μ peaks: 318/314	PNN50	4.70/5.46	8.55/10.42	+.967 [+.912, +.988]	+0.99	+0.94	+0.76 [-3.99, +5.50]
# seg: 17	LF	1740.17/1801.81	1293.81/1265.94	+.982 [+.952, +.993]	+0.98	+0.96	+61.63 [-416.36, +539.63]
	HF	603.85/747.19	934.21/1333.10	+.918 [+.792, +.969]	+0.98	+0.87	+143.34 [-754.77, +1041.46]
	LF:HF	5.26/7.53	2.90/6.65	+.625 [+.215, +.846]	+0.92	+0.49	+2.27 [-5.85, +10.38]
	LFnu	80.86/81.10	8.90/14.16	+.745 [+.420, +.900]	+0.81	+0.63	+0.24 [-16.67, +17.15]
	HFnu	19.14/18.90	8.90/14.16	+.745 [+.420, +.900]	+0.81	+0.63	-0.24 [-17.15, +16.67]
Rest	RMSS	18.85/17.93	8.86/9.97	+.935 [+.904, +.956]	+0.95	+0.88	-0.91 [-7.35, +5.53]
μ len: 249 s	SDNN	58.11/58.21	20.89/21.95	+.982 [+.974, +.988]	+0.98	+0.97	+0.10 [-7.89, +8.09]
μ peaks: 363/357	PNN50	3.09/3.14	4.66/5.39	+.954 [+.933, +.968]	+0.96	+0.92	+0.04 [-2.97, +3.06]
# seg: 111	LF	1324.54/1361.74	1051.67/1177.61	+.946 [+.946, +.946]	+0.95	+0.89	+37.20 [-724.00, +798.40]
	HF	372.58/365.69	519.14/656.80	+.914 [+.877, +.940]	+0.94	+0.86	-6.90 [-489.31, +475.52]
	LF:HF	5.16/6.47	2.96/3.63	+.614 [+.413, +.744]	+0.67	+0.30	+1.31 [-4.05, +6.67]
	LFnu	79.79/83.05	10.32/8.96	+.701 [+.525, +.808]	+0.75	+0.26	+3.26 [-10.40, +16.93]
	HFnu	20.21/16.95	10.32/8.96	+.701 [+.525, +.808]	+0.75	+0.26	-3.26 [-16.93, +10.40]
Bike (60 W)	RMSS	10.82/8.13	4.28/2.78	-.018 [-.292, +.303]	-0.02	-3.42	-2.70 [-12.81, +7.41]
μ len: 294 s	SDNN	29.20/27.11	9.68/9.07	+.793 [+.596, +.899]	+0.81	+0.53	-2.10 [-13.50, +9.31]
μ peaks: 489/483	PNN50	0.28/0.14	0.56/0.33	+.394 [+.045, +.662]	+0.46	-1.44	-0.14 [-1.12, +0.85]
# seg: 28	LF	247.06/240.91	185.05/182.23	+.945 [+.886, +.974]	+0.94	+0.88	-6.15 [-126.78, +114.49]
	HF	93.06/52.20	101.05/37.16	+.053 [-.271, +.389]	+0.09	-8.16	-40.86 [-245.64, +163.93]
	LF:HF	3.15/5.73	1.15/2.93	-.012 [-.208, +.250]	-0.03	-0.98	+2.59 [-3.65, +8.82]
	LFnu	73.41/81.02	9.83/11.49	-.065 [-.345, +.266]	-0.08	-1.32	+7.61 [-23.17, +38.40]
	HFnu	26.59/18.98	9.83/11.49	-.065 [-.345, +.266]	-0.08	-1.32	-7.61 [-38.40, +23.17]
Bike (120 W)	RMSS	13.07/6.91	5.13/3.01	+.064 [-.161, +.427]	+0.15	-6.94	-6.15 [-17.04, +4.73]
μ len: 272 s	SDNN	54.11/49.63	20.39/16.76	+.835 [+.552, +.946]	+0.87	+0.55	-4.48 [-24.52, +15.57]
μ peaks: 541/533	PNN50	0.52/0.18	0.63/0.27	+.355 [-.112, +.728]	+0.59	-4.47	-0.34 [-1.36, +0.67]
# seg: 13	LF	100.62/42.73	96.74/39.70	+.266 [-.168, +.669]	+0.47	-5.93	-57.90 [-225.22, +109.43]
	HF	114.36/32.75	112.56/33.92	-.136 [-.464, +.347]	-0.34	-19.56	-81.61 [-332.89, +169.66]
	LF:HF	1.07/3.49	0.61/3.89	+.053 [-.316, +.507]	+0.23	-0.37	+2.43 [-5.03, +9.88]
	LFnu	48.23/62.76	12.94/25.16	+.031 [-.386, +.512]	+0.05	-0.58	+14.53 [-39.90, +68.96]
	HFnu	51.77/37.24	12.94/25.16	+.060 [-.386, +.512]	+0.05	-0.58	-14.53 [-68.96, +39.90]
Walk	RMSS	14.67/8.93	3.54/2.46	+.144 [-.091, +.478]	+0.42	-6.66	-5.74 [-12.31, +0.83]
μ len: 291 s	SDNN	33.36/29.83	9.17/8.28	+.755 [+.358, +.910]	+0.81	+0.38	-3.54 [-14.15, +7.07]
μ peaks: 444/446	PNN50	1.00/0.20	1.46/0.31	-.008 [-.361, +.412]	-0.03	-30.34	-0.79 [-3.74, +2.15]
# seg: 17	LF	414.73/337.55	195.72/115.77	+.417 [-.017, +732]	+0.51	-1.59	-77.18 [-407.37, +253.02]
	HF	208.90/50.43	181.55/24.16	+.046 [-.193, +.386]	+0.30	-97.74	-158.47 [-503.17, +186.23]
	LF:HF	3.17/8.13	2.09/3.90	+.238 [-.104, +.614]	+0.63	-1.32	+4.95 [-1.01, +10.92]
	LFnu	69.65/87.26	15.27/4.73	+.128 [-.111, +.455]	+0.49	-22.01	17.61 [-9.04, +44.26]
	HFnu	30.35/12.74	15.27/4.73	+.128 [-.111, +.455]	+0.49	-22.01	-17.61 [-44.26, +9.04]

Holter follow similar trends, even though their level of agreement is low.

5.4 Monitoring during ambulatory conditions

Even though many studies, including ours, use the Holter monitor as a baseline, it may not be the best solution to monitor HR during strenuous sports. The Holter's cables and electrodes are very susceptible to movement. Therefore, wearable devices may provide better reliability under these conditions. Figure 9 shows an example of this case. We can observe that the Holter signal becomes noisy as the subject engages with the jog and running activities. However, this may not be the case when monitoring IBI. Figure 10 shows the Everion's and Holter's IBI of one subject during the rest and bike activities. Everion and Holter show good agreement during rest, but Everion shows more noise during the bike activity. Further experiments are required to determine if wearable devices can provide more reliability than a Holter monitor during sports activities.

6 CONCLUSIONS

We gather a dataset with multiple off-the-shelf wearable devices comprising several sensors used to track physiological data and made our dataset publicly available to the research community. We

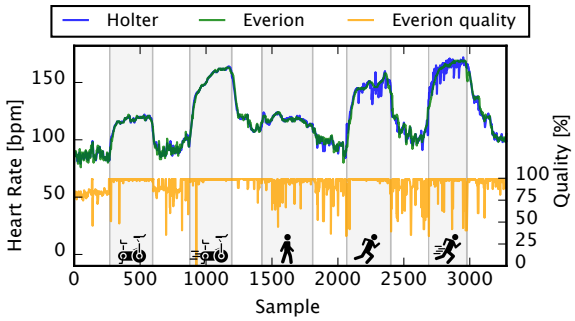

Figure 9: Signals collected using the Everion and medical-grade Holter monitor. In particular, the Everion device shows very good agreement with the data of the Holter monitor. The Holter shows less reliability (noise) during the jog and run activities.

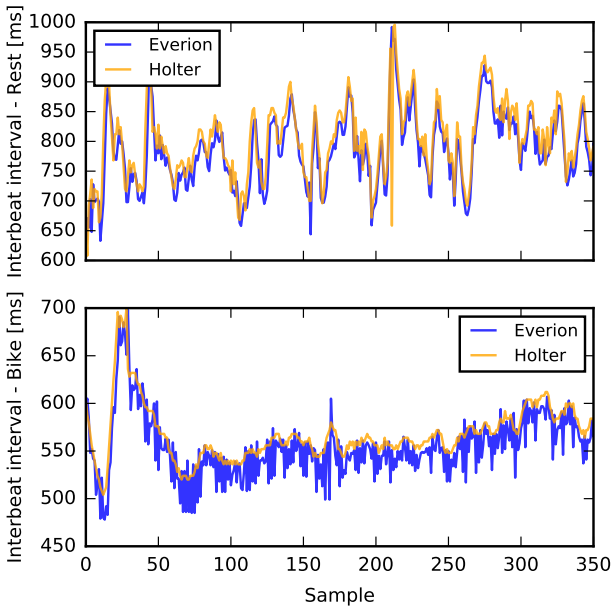

Figure 10: Comparing of Everion and Holter IBI during rest and bike activities. The devices show very good agreement during rest. Everion signals show more variation during the bike session.

focus on evaluating the agreement between mean HR and HRV metrics derived from the PPG sensor in wearable devices and a standard ECG Holter monitor under different physiological conditions. We show that armband-based devices dominate in precision when monitoring mean HR in all considered settings. Additionally, we show that the Everion device is a valid proxy for HRV metrics during periods not involving strenuous physical activity. Therefore, we hypothesize that the Everion is a potential candidate for continuous monitoring physiological data in persons with sedentary lifestyles such as office workers, patients, etc. Furthermore, we look into the future and challenge whether the Holter monitor is a better baseline than wearable devices for monitoring HR and HRV during strenuous activities and conclude that further exploration in this regard is needed. Finally, we show that participants have a preference for wrist-based devices and that their choices are not significantly affected by the predetermined duration of the monitoring. Thus, there is a trade-off between comfort and reliability when deciding between armband-based or wrist-based devices.

REFERENCES

[1] 1996. Report of the 1995 World Health Organization/International Society and Federation of Cardiology Task Force on the Definition and Classification of Cardiomyopathies. *Circulation* 93, 5 (1996), 841–842.

[2] NY: IBM Corp Armonk. Released 2017. IBM SPSS Statistics for Windows, Version 25.0.

[3] Raquel Bailón, Nuria Garatachea, Ignacio de la Iglesia, Jose Casajús, and Pablo Laguna. 2013. Influence of Running Stride Frequency in Heart Rate Variability Analysis During Treadmill Exercise Testing. *IEEE Transactions on Biomedical Engineering* 60, 7 (2013), 1796–1805.

[4] Liliana Barrios, Pietro Oldrati, Silvia Santini, and Andreas Lutterotti. 2018. Recognizing Digital Biomarkers for Fatigue Assessment in Patients with Multiple Sclerosis. In *PervasiveHealth*. EAI.

[5] Biovotion AG 2018. Everion monitor. Retrieved Nov 9, 2018 from http://www.biovotion.com/

[6] John Martin Bland and Douglas Altman. 1986. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. *The Lancet* 327 (1986), 307 – 310.

[7] Domenico Bonaduce, Mario Petretta, Fortunato Marciano, Maria Vicario, Claudio Apicella, Maria Rao, Emanuele Nicolai, and Massimo Volpe. 1999. Independent and incremental prognostic value of heart rate variability in patients with chronic heart failure. *American Heart Journal* 138, 2 (1999), 273 – 284.

[8] A. V. Challoner and C. A. Ramsay. 1974. A photoelectric plethysmograph for the measurement of cutaneous blood flow. *Physics in Medicine & Biology* 19, 3 (1974), 317–328.

[9] Tarani Chandola, Annie Britton, Eric Brunner, Harry Hemingway, Marek Malik, Meena Kumari, Ellena Badrick, Mika Kivimaki, and Michael Marmot. 2008. Work stress and coronary heart disease: what are the mechanisms? *European Heart Journal* 29, 5 (2008), 640–648.

[10] Empatica Inc. 2018. E4 wristband. Retrieved April 02, 2018 from https://www.empatica.com/en-eu/research/e4/

[11] Fitbit 2018. Fitbit Charge HR. Retrieved Nov 9, 2018 from https://www.fitbit.com/be/chargehr

[12] General Electric Healthcare 2018. Holter Recorder SEER* 1000. Retrieved Nov 9, 2018 from http://www3.gehealthcare.com/en/products/categories/diagnostic_ecg/ambulatory/seer_1000

[13] General Electric Healthcare 2019. CardioDay 2.5 Holter ECG. Retrieved Apr 19, 2019 from https://www.gehealthcare.com/en/products/diagnostic-ecg/ambulatory/cardioday-holter-ecg-software

[14] Konstantinos Georgiou, Andreas V. Larentzakis, Nehal N. Khamis, Ghadah I. Alsuhaibani, Yasser A. Alaska, and Elias J. Giallafos. 2018. Can Wearable Devices Accurately Measure Heart Rate Variability? A Systematic Review. *Folia Medica* 60, 1 (2018), 7 – 20.

[15] Nicholas Giardino, Paul Lehrer, and Robert Edelberg. 2002. Comparison of finger plethysmograph to ECG in the measurement of heart rate variability. *Psychophysiology* 39, 2 (2002), 246–53.

[16] David Giles, Nick Draper, and William Neil. 2016. Validity of the Polar V800 heart rate monitor to measure RR intervals at rest. *European Journal of Applied Physiology* 116, 3 (2016), 563–571.

[17] googlefitbit 2016. googlefitbit. Retrieved Nov 23, 2018 from https://github.com/simonbromberg/googlefitbit

[18] David Hernando, Nuria Garatachea, Rute Almeida, Jose Casajús, and Raquel Bailón. 2018. Validation of Heart Rate Monitor Polar RS800 for Heart Rate Variability Analysis During Exercise. *Journal of Strength and Conditioning Research* 32, 3 (March 2018), 716–725.

[19] Empatica Inc. 2018. How is IBI.csv obtained? Retrieved December 28, 2018 from https://support.empatica.com/hc/en-us/articles/201912319-How-is-IBI-csv-obtained-

[20] International Data Corporation 2018. IDC Forecasts. Retrieved Jan 08, 2018 from https://www.idc.com/getdoc.jsp?containerId=prUS44276818

[21] Riazul Islam, Daehan Kwak, Humaun Kabir, Mahmud Hossain, and Kyung-Sup Kwak. 2015. The Internet of Things for Health Care: A Comprehensive Survey. *IEEE Access* 3 (2015), 678–708.

[22] Edward Jo, Kiana Lewis, Dean Directo, Michael J Kim, and Brett A Dolezal. 2016. Validation of Biofeedback Wearables for Photoplethysmographic Heart Rate

Tracking. *Journal of sports science & medicine* 15, 3 (08 2016), 540–547.
[23] Andrew H. Kemp, Daniel S. Quintana, Marcus A. Gray, Kim L. Felmingham, Kerri Brown, and Justine M. Gatt. 2010. Impact of Depression and Antidepressant Treatment on Heart Rate Variability: A Review and Meta-Analysis. *Biological Psychiatry* 67, 11 (2010), 1067 – 1074. Synaptic Development in Mood Disorders.
[24] Paul Kligfield, Leonard S. Gettes, James J. Bailey, Rory Childers, Barbara J. Deal, E. William Hancock, Gerard van Herpen, Jan A. Kors, Peter Macfarlane, David M. Mirvis, Olle Pahlm, Pentti Rautaharju, and Galen S. Wagner. 2007. Recommendations for the Standardization and Interpretation of the Electrocardiogram. *Journal of the American College of Cardiology* 49, 10 (2007), 1109–1127.
[25] Terry Koo and Mae Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* 15, 2 (06 2016), 155–163.
[26] Hyunjae Lee, Tae Kyu Choi, Young Bum Lee, Hye Rim Cho, Roozbeh Ghaffari, Liu Wang, Hyung Jin Choi, Taek Dong Chung, Nanshu Lu, Taeghwan Hyeon, Seung Hong Choi, and Dae-Hyeong Kim. 2016. A graphene-based electrochemical device with thermoresponsive microneedles for diabetes monitoring and therapy. *Nature Nanotechnology* 11 (2016), 566–572.
[27] Sinziana Mazilu, Ulf Blanke, Michael Hardegger, Gerhard Tröster, Eran Gazit, and Jeffrey M. Hausdorff. 2014. GaitAssist: A Daily-life Support and Training System for Parkinson's Disease Patients with Freezing of Gait. In *Proc. CHI*. ACM Press, 2531–2540.
[28] Cameron McCarthy, Nikhilesh Pradhan, Calum Redpath, and Andy Adler. 2016. Validation of the Empatica E4 wristband. In *2016 IEEE EMBS International Student Conference (ISC)*. 1–4.
[29] Patrick McSharry, Gari Clifford, Lionel Tarassenko, and Leonard Smith. 2003. A dynamical model for generating synthetic electrocardiogram signals. *IEEE Transactions on Biomedical Engineering* 50, 3 (2003), 289–294.
[30] Scott Michael, Kenneth S Graham, and Oam Davis, Glen M. 2017. Cardiac Autonomic Responses during Exercise and Post-exercise Recovery Using Heart Rate Variability and Systolic Time Intervals-A Review. *Frontiers in physiology* 8 (05 2017), 301; 301–301.
[31] David Nunan, Djordje Jakovljevic, Gay Donovan, Lynette Hodges, Gavin Sandercock, and David Brodie. 2008. Levels of agreement for RR intervals and short-term heart rate variability obtained from the Polar S810 and an alternative system. *European Journal of Applied Physiology* 103, 5 (2008), 529–537.
[32] Jakub Parak and Ilkka Korhonen. 2014. Evaluation of wearable consumer heart rate monitors based on photopletysmography. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 3670–3673.
[33] Mitesh Patel, Sara Lal, Diarmuid Kavanagh, and Peter Rossiter. 2011. Applying neural network analysis on heart rate variability data to assess driver fatigue. *Expert Systems with Applications* 38, 6 (2011), 7235 – 7242.
[34] Polar 2018. Polar OH1. Retrieved Nov 9, 2018 from https://www.polar.com/en/products/accessories/oh1-optical-heart-rate-sensor
[35] Axel Schäfer and Jan Vagedes. 2013. How accurate is pulse rate variability as an estimate of heart rate variability? *International Journal of Cardiology* 166, 1 (jun 2013), 15–29.
[36] Fredric Shaffer and J P Ginsberg. 2017. An Overview of Heart Rate Variability Metrics and Norms. *Frontiers in Public Health* 5 (09 2017), 258.
[37] Julian Thayer and Richard Lane. 2007. The role of vagal function in the risk for cardiovascular disease and mortality. *Biological Psychology* 74, 2 (2007), 224 – 242.
[38] Robert Thiebaud, Merrill Funk, Jacelyn Patton, Brook Massey, Terri Shay, Martin Schmidt, and Nicolas Giovannitti. 2018. Validity of wrist-worn consumer products to measure heart rate and energy expenditure. *Digital health* 4 (04 2018).
[39] Basilio Vescio, Maria Salsone, Antonio Gambardella, and Aldo Quattrone. 2018. Comparison between Electrocardiographic and Earlobe Pulse Photoplethysmographic Detection for Evaluating Heart Rate Variability in Healthy Subjects in Short- and Long-Term Recordings. *Sensors (Basel, Switzerland)* 18, 3 (2018).
[40] Wahoo Fitness 2018. Wahoo Ticker Fit. Retrieved Nov 9, 2018 from https://eu.wahoofitness.com/devices/heart-rate-monitors
[41] Matthew Wallen, Sjaan Gomersall, Shelley Keating, Ulrik Wisløff, and Jeff Coombes . 2016. Accuracy of Heart Rate Watches: Implications for Weight Management. *PLOS ONE* 11, 5 (2016).