International Journal of Psychophysiology 182 (2022) 105–118

Contents lists available at ScienceDirect

International Journal of Psychophysiology

journal homepage: www.elsevier.com/locate/ijpsycho

Validity of the Empatica E4 wristband to estimate resting-state heart rate variability in a lab-based context

Hans Stuyck [a,b,*], Leonardo Dalla Costa [a], Axel Cleeremans [b], Eva Van den Bussche [a]

[a] *KU Leuven, Faculty of Psychology and Educational Sciences, Brain & Cognition, Tiensestraat 102, 3000 Leuven, Belgium*
[b] *Université libre de Bruxelles, Faculty of Psychology and Education Sciences, Center for Research in Cognition and Neurosciences, Franklin Rooseveltlaan 50, 1050 Brussel, Belgium*

ARTICLE INFO

Keywords:
Empatica E4
Validity
Ultra-short intervals
ECG
Heart rate
Heart rate variability

ABSTRACT

Lab research might benefit from the advantages of wearable devices, such as their ease of use, to estimate pulse rate (PR) and pulse rate variability (PRV) as an equivalent for heart rate (HR) and heart rate variability. However, before implementing them in a lab context, the validity of the PR and PRV, also on ultra-short time scales (e.g., 30s), needs to be confirmed. We recorded heart activity simultaneously with an E4 wristband and an ECG device in a seated resting condition for 5 min. Our results showed that HR, RMSSD, SDNN and LF, but not HF, were validly estimated by the E4 wristband. Furthermore, the E4 wristband could validly estimate PR with recording lengths as short as 10s. RMSSD and SDNN were validly estimated using 30s or 120 s or an average of multiple short intervals (10s), while HF likely requires longer recording intervals. Based on this study, we formulated several recommendations for using the E4 wristband in a lab context.

1. Introduction

The study of heart rate variability (HRV) has been the object of scientific scrutiny for centuries (for a historical overview, see Berntson et al., 1997). HRV relates to the variation in beat-to-beat intervals (i.e., interbeat interval; IBI). These fluctuations of the IBI result from the active interplay between the parasympathetic (i.e., PNS; rest and digest) and the sympathetic (i.e., SNS; fight or flight) nervous system (Shaffer et al., 2014). In general, it has been found that higher HRV is associated with better adaptability to the environment, for example, concerning physical health, stress regulation, and self-regulation (Graham et al., 2019; Ottaviani et al., 2018; Pulopulos et al., 2018). The most commonly used HRV metrics are situated in the time and frequency domain (Shaffer and Ginsberg, 2017). Time-domain metrics represent the variability of the IBI. For example, the standard deviation of normal IBIs (SDNN) and the root mean square of successive differences between normal IBIs (RMSSD). Frequency domain metrics are represented by the absolute (relative or normalized) power of a signal in one of the four frequency bands related to HRV (i.e., ultra low frequency [ULF; $\leq$0.003 Hz], very low frequency [VLF; 0.0033–0.04 HZ], low frequency [LF; 0.04–0.15 Hz] and high frequency [HF; 0.15–0.4 Hz]; Shaffer and Ginsberg, 2017). The informative value of HRV for the individual's mental and physical adaptability has given rise to its use in research settings (see Ottaviani et al., 2018; Stone et al., 2021). HRV is typically measured in laboratory settings by connecting participants to an electrocardiogram device (i.e., ECG). The ECG represents the heart's bioelectrical activity through several waves, of which the large r-peak is used to determine the IBI (Silverthorn, 2004). The ECG's fine-grained representation of the heart's activity makes it the current gold standard to measure HRV (Berntson et al., 1997). However, an ECG device is relatively expensive. Hence, labs often only have one ECG device so that just one participant at a time can be tested. This increases the cost of using ECG in terms of labor cost and meeting sample size requirements. Although ECG studies have been performed on diverse populations (e.g., Ajayi et al., 2021; Nuske et al., 2021), we contend that some targeted populations in lab-based studies might benefit from easier, less time-consuming, and less invasive heart recording setups. For example, young children, older adults, or patients who are already regularly submitted to testing might benefit from a simpler, less obtrusive setup.

One promising solution could be the use of wearable devices in the lab (e.g., Polar smartwatches). These wearables have been developed for daily use, informing users about the complex interaction between behavior, environment, and physical health (Castaneda et al., 2018; Ishaque et al., 2021). These wearable devices are considerably cheaper

* Corresponding author at: KU Leuven, Faculty of Psychology and Educational Sciences, Brain & Cognition, Tiensestraat 102, 3000 Leuven, Belgium.
E-mail addresses: hans.stuyck@kuleuven.be (H. Stuyck), axcleer@ulb.ac.be (A. Cleeremans), eva.vandenbussche@kuleuven.be (E. Van den Bussche).

https://doi.org/10.1016/j.ijpsycho.2022.10.003
Received 1 July 2022; Received in revised form 10 October 2022; Accepted 11 October 2022
Available online 14 October 2022
0167-8760/

than an ECG device so that multiple wearables can be bought for the price of one ECG device. Moreover, their design as a watch makes them an almost unnoticeable device that does not hamper its users (Castaneda et al., 2018). However, the estimation of HRV by such devices is quite different from its ECG-based estimation. It is based on the changes in blood volume measured in the body's periphery (e.g., wrist). The differences in blood volume result from the heart contracting to push blood out (i.e., systole) and from the heart's subsequent relaxation (i.e., diastole; Shaffer et al., 2014). These differences in blood volume are measured via photo-plethysmography (PPG). This technique measures the amount of light absorbed by the blood vessels, which is proportional to the variations in blood volume. These blood volume variations are represented by the systolic and diastolic pulse waves jointly representing the blood-volume-pulse signal (i.e., BVP; see Fig. 1; Alqaraawi et al., 2016). The time interval between the fiducial point on each successive systolic pulse wave is used to estimate the pulse-to-pulse interval (i.e., PPI) on which pulse rate (PR; i.e., beats per minute) and pulse rate variability (i.e., PRV) are calculated, as an equivalent to assessing heart rate (HR) and HRV based on the r-peaks of the ECG (Alqaraawi et al., 2016).

However, there exists a time lag between the depolarization of the ventricles (r-peak) initiating the mass ventricular contraction of the heart and the increase of blood volume measured in the body's periphery (i.e., the pulse transit time). Features of the blood vessels influence this time lag (e.g., elasticity, vascular diameter) that can change the shape of the pulse waves, resulting in the misalignment between the r-peak and the systolic pulse wave. This might affect the temporal dynamics of the PPI and, thus, the accuracy of the HRV estimate since alterations of the PPI dynamics would not alter PRV but rather the difference between PRV and HRV (Lu and Yang, 2009). Furthermore, several other factors might introduce errors in the accuracy of the PRV measurement. For example, the sampling frequency of such wearable devices is considerably lower than an ECG device (e.g., Empatica E4 wristband = 64 Hz). This lower sampling frequency provides a less fine-grained temporal resolution representing both pulse waves (Laborde et al., 2017). A clear and distinct pulse wave morphology is critical to accurately detect the fiducial point on the systolic wave. Also, the PPG-based PRV estimation is more vulnerable to orthostatic changes, motion artifacts, stress induction, and respiratory patterns than ECG-based HRV (see Mejía-Mejía et al., 2020 for a review).

So before implementing such wearable devices in lab research, it is vital to assess how valid the PRV metrics derived by such devices are (Schuurmans et al., 2020). The most common way to assess the validity of PPG-based PR and PRV metrics as approximations of HR and HRV metrics obtained by wearable devices is by comparing them to a gold standard reference (Schuurmans et al., 2020). HR and HRV metrics obtained by an ECG device are often proposed as such a gold standard reference (e.g., van Lier et al., 2020).

1.1. Validity of E4 wristband-based PRV

One wearable device that its manufacturers termed fit for lab-based studies is the Empatica E4 wristband (i.e., E4 wristband; https://www.empatica.com/en-eu/). The E4 wristband has already been used to study the role of PR and/or PRV in a variety of lab-based research topics, such as stress (e.g., Mishra et al., 2020), music therapy (e.g., Rahman et al., 2021), and emotions (e.g., So et al., 2021).

To date, only a limited number of studies have examined the *validity* of the PR and PRV metrics obtained by the E4 wristband. For instance, Menghini et al. (2019) examined the validity of the E4 wristband's PR and PRV metrics as accurate approximations of ECG-based HR and HRV metrics under various conditions (i.e., during seated position, paced breathing, orthostatic posture, slow walking, keyboard typing, Stroop test, speech preparation, public speech, and speech recovery). They showed that PR obtained by the E4 wristband and HR acquired with the gold standard ECG device was comparable across conditions. The PRV and HRV metrics in the time (SDNN and RMSSD) and frequency domain (LF and HF) were comparable during seated and paced breathing conditions, but these E4 wristband-based PRV metrics were less accurate approximations of the ECG-based HRV metrics during conditions that involved movement (i.e., slow walking) or cognitive performance (i.e., Stroop test). The authors demonstrated that this decrease in PRV accuracy in those conditions was likely due to hand/wrist movement (see Ryan et al., 2019, van Lier et al., 2020, for similar results). Furthermore, in contrast to the studies mentioned above, Schuurmans et al. (2020) observed that E4 wristband PRV-based RMSSD was not comparable to the ECG-based RMSSD during a seated-rest condition (see Ollander et al., 2016 for similar results). Additionally, McCarthy et al. (2016) simultaneously measured ECG using a portable ECG device and BVP using the E4 wristband during a 24 h or 48 h recording. They showed that in most cases (85 %), signal quality assessed based on visual inspection (shape, stability, and noise) was comparable between the E4 wristband and the ECG device, although this was only the case during the less-active parts of the day (i.e., night and morning). Hence, previous studies assessing the validity of the PR and PRV metrics obtained with the E4 wristband seem to indicate that in seated or low activity

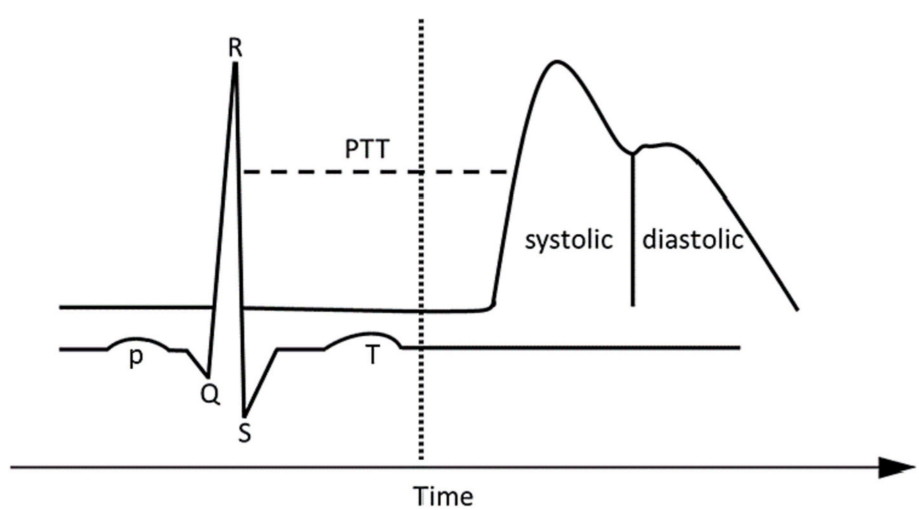

Fig. 1. An example of an ECG signal and BVP signal.
Note. On the left side of the vertical dotted line, the bio-electrical signal of the heart is shown, and on the right side, the systolic and diastolic pulse waves; PTT: pulse transit time.

conditions, the E4 wristband can provide comparable PR and PRV metrics as the HR and HRV metrics acquired with a gold standard ECG device, although for some indices (e.g., RMSSD) contradictory results have been obtained. However, inconsistencies in the statistical procedures used to assess the E4 wristband's validity (e.g., difference factor, visual inspection, and/or Bland-Altman plots; McCarthy et al., 2016; Ollander et al., 2016; Schuurmans et al., 2020), as well as the often small sample sizes (e.g., $N = 7$; Ollander et al., 2016), make it currently difficult to draw strong conclusions.

1.2. Validity of E4 wristband-based PRV obtained with ultra-short-term intervals

As a gold standard, 5-min intervals are stated as sufficient to measure PR/HR and PRV/HRV reliably (Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology, 1996; from now on Task Force, 1996). However, lab experiments often consist of multiple short-interval events (e.g., "trials") occurring within one testing session (i.e., intervals <5 min; e.g., Choi et al., 2017; Lackner et al., 2013; Shen et al., 2017). A limited number of ECG studies have already compared the HR and HRV metrics obtained with UST intervals to those acquired with the gold standard 5 min interval (e.g., Baek et al., 2015). For example, Shaffer et al. (2016) measured HR and HRV metrics via ECG with UST intervals (i.e., interval length = 10s, 20s, 30s, 60s, 90s, 120 s, 180 s, and 240 s) in a seated-rest condition. They subsequently compared those HR and HRV metrics obtained with the UST intervals to those acquired with a 5 min interval. They found that only HR across all UST intervals was highly correlated with the 5 min HR. RMSSD and SDNN needed at least a 60s UST interval, and LF and HF power were only validly estimated with at least the 90s and 180 s UST intervals, respectively (Shaffer and Ginsberg, 2017; Shaffer et al., 2016; see also Baek et al., 2015). Munoz et al. (2015), on the other hand, showed that RMSSD and SDNN might already be validly estimated with a UST interval of 10s (see also Nussinovitch et al., 2011). However, to our knowledge, such a comparison between UST intervals and the gold standard 5 min interval has never been studied for the E4 wristband's PR and PRV metrics. This is particularly important to establish the E4 wristbands' applicability as a PR and PRV recording device in lab studies comprising multiple fast trials.

Given this overview, the aim of the current study was twofold. First, we aimed to evaluate the validity of the E4 wristband's PR and PRV metrics as accurate approximations of HR and HRV by comparing it to a gold standard ECG device in a lab context. Second, we aimed to examine the validity of the E4 wristband's UST interval measurement of PR and PRV to assess their usefulness for implementation in lab-based studies with short-interval events. We solely focused on a seated-rest condition in a lab context.

2. Method

2.1. Participants

A convenience sample of 79 undergraduates of the KU Leuven participated in this study in return for course credits. This sample was recruited in the context of a broader study on creative problem-solving (Stuyck et al., in preparation). In line with Quintana and Heathers (2014), several inclusion criteria (i.e., nonsmokers, body-mass index [BMI] <30, Beck depression inventory score < 29, no cardiovascular or neurological medication use, no history of or current cardiopulmonary diseases, psychiatric disorders, and/or neurological disorders) and instructions concerning daily routines immediately preceding participation (i.e., no alcohol consumption the night before and day of the experiment, at least 6 h sleep the night before the experiment, no caffeine consumption during the 2 h before the experiment, no heavy meal consumption and no strenuous physical activity prior to the experiment) were used, as they might negatively influence the BVP and ECG recording.

We excluded seven participants based on technical issues with the ECG signal (i.e., incorrect electrode placement, disconnection of the electrodes during recording, or trigger interface unresponsiveness). In line with van Lier et al. (2020), we visually inspected the quality of the BVP signal morphology. We excluded participants if >20 % of the recording interval contained an unstable BVP signal, indicating unusable data (see Appendix A for supplementary materials where supplementary material A depicts an example). This led to the exclusion of 26 participants. We also visually inspected the quality of the ECG signal morphology for abnormalities and stability (see supplementary material A for an example). This led to the additional exclusion of five participants (see Kumral et al., 2019; Shaffer and Ginsberg, 2017). After that, we identified outlying HRV observations by using the Tukey method (1997) of the median ± three times the interquartile range (see Kumral et al., 2019 for a similar procedure). Based on this method, we excluded two participants with outlying PRV/HRV observations on two or more different PRV/HRV indices. Finally, we identified two participants with only outlying BVP-based and ECG-based LF. Therefore, we decided to exclude these participants from the statistical analyses involving LF solely.[1] This resulted in a final sample of 39 participants (mean age = 19, $SD = 1.55$, range = 17–24, 35 female). Our sample size was based on van Lier et al. (2020), who performed an a priori sample size calculation. Based on the large effect sizes observed in previous studies comparing the E4 wristband to a gold standard ECG device (i.e., minimum effect size of $r = 0.72$; Schuurmans et al., 2020), and when comparing UST HRV to 5 min recording intervals (i.e., minimum effect size of $r = 0.76$; Munoz et al., 2015), our study with a sample of $N = 39$ had an estimated power of 0.99 to detect such large effects (Campbell and Thompson, 2012). Before the start of the experiment, all participants provided written informed consent. The social and societal ethics committee of the KU Leuven approved this study (approval code G-2019 121,929).

2.2. Assessment and measurement

Before assessing the PRV and HRV indices crucial for this study, we note that participants completed six practice trials of a cognitive task, namely the compound remote associates test (CRAT; see https://osf.io/snb3k/). Once the PRV and HRV measurements relevant to this study were collected, participants performed this cognitive task as part of a broader study. As this task was not relevant to the current research questions and was only assessed *after* the critical PRV and HRV indices were collected, it will not be discussed further (for a detailed description of the CRAT instructions and experimental procedure, see https://osf.io/4frcb/).

2.3. Equipment

Participants were seated individually in a quiet, dimly lit room held at a constant temperature between 21° and 23°. They faced the computer monitor from approximately 60 cm. A Dell Optiplex 3060 computer was used with a Dell 23.6-in. monitor.

2.3.1. Nexus-10 MKII ECG device

Nexus-10 MKII (Mind Media BV, Herten, the Netherlands) was used as the gold standard ECG recording device (CE-certified; 93/42/EC Annex XII). The device obtains the ECG signal in microvolts with a sampling rate of 256 Hz. Three pre-gelled Ag/AgCl electrodes were used. Following the modified Lead-II placement, these were attached to the upper body (see Kuipers et al., 2017). Namely, the negative electrode

[1] We also performed the data analysis with the outliers included. The results obtained with this analysis are similar to the results of the main analysis described in the main text. This data analysis can be found in supplementary material B.

below the center of the right collarbone, the positive electrode on the lower left rib cage and the ground electrode below the left collarbone. Before placing the electrodes, the skin was cleaned with an alcohol pad.

2.3.2. Empatica E4 wristband

The Empatica E4 wristband (Empatica, Milan, Italy) is a certified medical wrist-worn device (CE-certified. No. 1876/MDD) that enables real-time multi-sensor data acquisition. The device allows the recording of four psychophysiological indices: BVP, acceleration, skin conductance, and skin temperature. The current study only used the BVP. The E4 wristband extracts BVP via a PPG sensor with two green and red photodiodes (LEDs). This BVP signal is acquired at a sampling rate of 64 Hz. Following the manufacturer's guidelines, participants wore the E4 wristband on their non-dominant hand to diminish the likelihood of motion artifacts.

2.4. Procedure

All participants were assessed between 9 am and 5 pm. Before entering the laboratory, the experimenter stressed that it was important to go to the toilet before testing if needed, as this might influence PRV/HRV data. After that, participants signed the informed consent. Their eligibility was assessed based on the inclusion criteria and adherence to the instructed daily routines. Participants were then given instructions on how to attach the ECG device's electrodes, which was accompanied by an image depicting the exact modified Lead-II placement. They also attached the E4 wristband to the wrist of their non-dominant hand. The experimenter visually checked if both devices were attached correctly. Participants then took place at the test computer. The experimenter explained that it was important to remain in an upright seated position without crossing their legs and to minimize sudden movements as much as possible (e.g., arm stretching, excessively coughing). It was explained that deviating from these instructions might cause BVP and ECG data acquisition issues. Participants then completed the six practice trials of the CRAT computer task. The experimenter explained that there would be a 10 min resting period after these practice trials. During this 10 min resting period, participants were instructed to relax and control their emotions to minimize the likelihood of engaging in ruminative or emotionally valenced thoughts that might affect the PRV/HRV recording. To minimize recording-awareness-related changes in their behavior, we told participants that we were mainly interested in their (later) performance during the cognitive task. This 10 min period consisted of a 5 min acclimatization period and a subsequent 5 min baseline period. It is this 5 min baseline period that was used in this study to validate the E4 wristband. After the 10 min resting period, participants completed the CRAT.

2.5. Data preprocessing

The E4 wristband and ECG device data were synchronized by relying on the computer's clock time (see Menghini et al., 2019; Milstein and Gordon, 2020). The Empatica software responsible for the E4 wristband recording synchronizes to the computer's clock time. Therefore, we created a timestamp of the computer's clock time while at the same time sending a trigger to the ECG device at the onset of the experiment. Subsequently, we matched the clock time provided by this timestamp, represented by the trigger in the ECG device, to that of the clock time registered by the Empatica software, thereby creating a synchronous time point of reference for the ECG device (via the trigger) and the E4 wristband. To ensure tight time synchronization between the two devices, we visually double-checked the concordance between the PPI and RRI time series.

Both the data obtained by the ECG device (i.e., ECG signal) and the E4 wristband (i.e., BVP signal) were preprocessed using Kubios premium (v. 3.4.2; Tarvainen et al., 2014 and see Tarvainen et al., 2020). The IBIs were calculated by determining the time interval between two r-peaks for the ECG signal and between two fiducial points on the systolic pulse wave for the BVP signal. To correct potential artifacts in the IBI time series, we used the automatic artifact correction algorithm of Kubios (see Lipponen and Tarvainen, 2019). All detected artifacts are subsequently replaced with IBIs based on cubic spline interpolation. To accommodate the non-stationarity of the IBI time series, Kubios deploys a detrending procedure by defining an a priori smoothing parameter (cut-off frequency 0.035 Hz; see Tarvainen et al., 2002). Additionally, all ECG and BVP signals were visually inspected for abnormal signals, unstable recording epochs in the recording interval, missed p-waves/r-peaks and missed artifacts by the algorithms that might influence the PRV/HRV data (e.g., supraventricular extrasystole; see Kumral et al., 2019 for a similar procedure). In case of any ECG or BVP signal abnormalities missed by the algorithms, we applied a manual correction to the signal (e.g., marking an unstable signal epoch as a to-be-excluded noise segment or adding missed p-waves/r-peaks; see supplementary material C for an overview of the percentage of corrected IBIs and noise-free ECG/BVP recording intervals).

2.5.1. Validity of E4 wristband-based PRV

Besides mean PR and HR (i.e., expressed in beats per minute; bpm), we used several PRV and HRV metrics to compare both recording devices. From the time domain, RMSSD and SDNN (both expressed in ms) were used and, from the frequency domain, the absolute power in the LF and HF bands (both expressed in ms^2) were used (similar to Menghini et al., 2019; Schuurmans et al., 2020; van Lier et al., 2020). We did not include the LF/HF ratio, as it remains an ambiguous HRV metric (Heathers, 2014).

2.5.2. Validity of E4 wristband-based PRV obtained with UST intervals

To compare the E4 wristband's PR and PRV obtained with the UST intervals to those same metrics acquired with the 5 min interval, we used RMSSD, SDNN, and HF (similar to Baek et al., 2015; Munoz et al., 2015; Salahuddin et al., 2007). The estimation of power in the HF band requires at least 10 oscillations (i.e., 70 s; Task Force, 1996). As BVP-based PRV estimation is considered less stable than ECG-based HRV estimation, we argue that it is unlikely that this bare minimum of 70s will suffice for an accurate estimation of HF. Therefore, we chose only to use a UST interval of 120 s to estimate HF. We did not assess LF as it requires at least 10 oscillations (i.e., 250 s), making it impossible to estimate it with UST HRV recordings (Pecchia et al., 2018; Task Force, 1996). All six UST intervals were segmented from the 5 min interval. We randomly extracted three nonoverlapping 10s segments from the 5 min recording for each participant. Note that the extraction was nonsequential, in the sense that T1 did not necessarily precede T2, etc. These 10s segments were used independently to assess the PR and PRV metrics and calculate an average value of the PR and PRV metric across those three 10s segments. As the 10s intervals only contain very few IBIs, we only accepted a 100 % stable BVP signal without corrected IBIs. If this was not the case, we rejected the 10s interval and randomly selected a new one. This process we repeated until we obtained three clean, valid 10s intervals. The UST interval of 30s existed of the first 30s of the 5 min interval, and the 120 s UST interval existed of the 120 s interval after those first 30s. This ensures their independence in validly estimating PR and PRV (see Munoz et al., 2015 for a similar procedure).

Our first UST analysis assessed the E4 wristband's internal consistency concerning PR and PRV measured at UST intervals. This is vital as it shows that measuring PR and/or PRV at these time scales can be surrogates for their 5 min estimation with the same device, thereby taking variability specific to BVP into account. However, the ECG 5 min recording can still be considered as the best approximation of the genuine mean HR and HRV. Therefore, in a second UST analysis, we compared the PR and PRV estimation of the UST intervals to the 5 min ECG-based estimation of HR and HRV. In the Results section, we summarized the results from this second UST analysis, and full details can be found in supplementary material D.

2.6. Statistical analysis

A *three-step hierarchical procedure* was used to assess the validity of PRV metrics obtained with the E4 wristband and UST recording intervals (Pecchia et al., 2018; Shaffer et al., 2020; van Lier et al., 2020).

2.6.1. Step 1, PRV metric selection

A Pearson product-moment correlation coefficient was used to examine the association strength between the proxies (PRV metrics obtained with the E4 wristband or UST interval) and the gold standard measurement (HRV metrics acquired with the ECG device or PRV metrics acquired with the 5 min interval). We followed the procedure of Menghini et al. (2019) and the recommendations of Pecchia et al. (2018). We used a correlation coefficient cut-off of $r = 0.70$ to identify the viable proxies of the gold standard measurement. Only the proxies that showed a correlation with the gold standard greater than or equal to $r = 0.70$ were retained for further analysis.

2.6.2. Step 2, PRV metric validity

Subsequently, a Bland-Altman plot was created for the viable proxies selected in step 1. Here, the differences between the proxy and the gold standard measurement are plotted against the gold standard measurement (see Giavarina, 2015; Menghini et al., 2021, for an in-depth explanation). We plotted against this gold standard instead of against the mean of the proxy and the gold standard, as suggested by Bland and Altman (1995), because in the current study, the gold standard is expected to be the best approximation of the genuine PRV and HRV. Therefore, it is also expected to have lower error variance and bias and, thus, to be better suited than the mean of the proxy and gold standard as a reference to plot against (see Krouwer, 2008; Munoz et al., 2015, for similar argumentation).

First, the mean of this difference was analyzed. This mean reflects the bias in measurement, as a perfect agreement between the proxy and the gold standard measurement would be reflected by a mean difference of zero. This bias can reflect a tendency of E4 wristband or UST interval to over/underestimate the proxy relative to the gold standard. The 95 % confidence intervals of the bias (i.e., 95 % CI) were used to assess this (Menghini et al., 2021). If zero is below/above the lower/upper bounds of this 95 % CI, there is an indication of a tendency to over/underestimate the proxy relative to the gold standard.

Second, it was measured to what extent the observed differences between the proxy and the gold standard were within acceptable limits of agreement. The limits of agreement (LOA) are represented by the mean of the differences (bias) $\pm$ 1.96*SD* (Menghini et al., 2021). The bounds of these LOA mark the inclusion of 95 % of the observed differences (i.e., 95 % LOA). To measure if those 95 % LOA can be considered acceptable limits in which the majority of the differences between the estimated proxy and gold standard lie, a priori LOA need to be determined before constructing the Bland-Altman plot (Giavarina, 2015). If the LOA are within the a priori LOA, the bounds marking the inclusion of 95 % of the observed differences are within the limits of a maximum acceptable deviation of the gold standard. This would imply that the proxy sufficiently agrees with the gold standard.

The determination of the a priori LOA requires the consideration of several features of the research (Giavarina, 2015; van Lier et al., 2020). For instance, the clinical necessity (e.g., the aimed-for diagnostic accuracy of the proxy), biological features of the sample being studied, the aim of the study, the time interval used to estimate PR and HR, and PRV and HRV (e.g., seconds, minutes or hours), and the parameters being studied (e.g., HR and RMSSD; van Lier et al., 2020). In the literature, mainly two a priori LOA have been proposed: an a priori LOA of 150 % (e.g., Menghini et al., 2019) and an a priori LOA of 110 % (e.g., van Lier et al., 2020). For example, an average 50 ms RMSSD measured with the gold standard leads to an a priori 150 % LOA of $\pm$25 ms (i.e., *lower bound* $= \frac{\bar{x}\text{ gold standard}*150}{100} - \bar{x}$ gold standard; upper bound $= \frac{\bar{x}\text{ gold standard}*150}{100}$). Thus, the 95 % LOA should lie within these limits of the maximum acceptable deviation of the gold standard of $\pm$25 ms. Although an a priori LOA of 110 % is considerably stricter than one of 150 %, this stricter a priori LOA is generally used to assess the utility of medical equipment (Advancement of Medical Instrumentation, 2002; van Lier et al., 2020). In this case, such a strict a priori LOA seems justified as medical diagnoses should depend on highly accurate derived metrics. However, as the current study aimed to validate the proxies obtained with E4 wristband and UST intervals within a lab research context with a maximal recording length of 5 min, we argue that the margin of error between the proxy and gold standard can be less strict. Therefore, we used an a priori LOA of 150 % calculated as in the example above, consistent with similar previous studies (Charlot et al., 2009; Menghini et al., 2019; Pichon et al., 2006).

Several assumptions need to be considered for the construction of the Bland-Altman plot. The bias and the 95 % LOA in the Bland-Altman plot are crucial for its interpretation. However, their accurate representation also depends on the absence of a positive/negative association between the observed differences and the gold standard value (i.e., proportional bias), an equal spread of the observed differences for each gold standard value (i.e., homoscedasticity), and the normal distribution of the observed differences between the proxy and the gold standard (i.e., normality). If one of these assumptions is not met, the above-explained representations of the bias and 95 % LOA no longer hold. In that case, the bias and 95 % LOA need to be represented via other calculations taking into account the violations (see supplementary material E for the precise calculations). To identify and accommodate any violations of the assumptions in our data, we followed the procedure as described by Altman and Bland (1983), Bland and Altman (1999), Euser et al. (2008), and Menghini et al. (2021). Regardless of the type of calculation used to represent the bias and its 95 % LOA, we always represented the observed differences, the gold standard values, the bias, the 95 % LOA, and the a priori 150 % LOA on their original scale to enhance interpretability.

2.6.3. Step 3, magnitude of difference

In the final step of the procedure, we calculated Cliff's delta (d) effect size to estimate the degree of overlap between the distributions of the proxy and the gold standard. This nonparametric effect size is less vulnerable to skewed, non-normal, and heteroscedastic data than, for example, Cohen's d (Romano et al., 2006). Its absolute values range from 0 to 1. We interpreted Cliff's d effect size as follows: $d \leq 0.15$ = negligible, $0.15 < d \leq 0.33$ = small, $0.33 < d \leq 0.5$ = medium, and $d > 0.5$ = large (see Romano et al., 2006).

We used the open-source *R* language and environment to perform statistical analysis (R Core Team, 2021). For the Bland-Altman plots, we used the source code provided by Menghini et al. (2021). We adjusted this source code to the specifics of the current study. To compute Cliff's d and its corresponding 95 % confidence intervals, we used the "effsize" package (Torchiano, 2020). All R code can be found on the Open Science Framework (https://osf.io/4frcb/).

Table 1
Correlation coefficients and the descriptive statistics of PR/HR and PRV/HRV metrics obtained with E4 and ECG.

Metrics	*r*[95 % CI]	*M*(*SD*) E4	*M*(*SD*) ECG
PR/HR	0.9989[0.9982, 0.9997]	82.06(7.13)	82.20(7.10)
RMSSD	0.92[0.88, 0.97]	39.33(10.65)	33.34(11.43)
SDNN	0.98[0.97, 1.00]	44.50(13.29)	42.92(13.91)
LF	0.96[0.94, 0.99]	1030.22(674.02)	1066.43(704.99)
HF	0.95[0.91, 1.00]	764.77(518.23)	637.73(450.89)

Note. PR, mean pulse rate (bpm); HR, mean heart rate (bpm); RMSSD, root mean square of successive differences between normal IBIs (ms); SDNN, standard deviation of normal IBIs (ms); LF, low frequency (ms^2); HF, high frequency (ms^2); *r,* Pearson correlation coefficient; 95 % CI, 95 % confidence intervals.

3. Results

3.1. Validity of E4 wristband-based PRV

3.1.1. Step 1, PRV metric selection

Table 1 depicts the correlation coefficients and the descriptive statistics of the mean PR and HR, the time (RMSSD and SDNN), and frequency (LF and HF) domain PRV and HRV metrics. Correlations between both devices surpassing the cut-off of $r = 0.70$ were found for HR, RMSSD, SDNN, LF, and HF. Therefore, the mean PR and all PRV metrics were included in step 2 of the three-step hierarchical procedure.

3.1.2. Step 2, PRV metric validity

Table 2 presents the results, and Fig. 2 presents the Bland-Altman plots (see supplementary material E for the assumptions handling). Concerning mean PR/HR, there was a slight tendency of the E4 wristband to underestimate it relative to its estimation by the ECG device. Namely, the upper bound of the bias's 95 % CI (i.e., −0.006) was just below zero. However, the 95 % LOA was within the a priori 150 % LOA, with 100 % of the observed differences lying within the a priori 150 % LOA bounds. These findings illustrated that the deviations between the mean PR obtained with the E4 wristband and the mean HR acquired with the ECG device were within the maximum acceptable deviation limits.

For RMSSD, the 95 % CI of its bias was entirely above zero, indicating that the E4 wristband tended to overestimate RMSSD relative to the same metric acquired with the ECG device. For SDNN, on the other hand, proportional bias was detected (i.e., a negative association between the observed differences and the ECG-device's estimation of SDNN; see Fig. 2). Specifically, SDNN was overestimated by the E4 wristband for observations of ECG-based SDNN lower than 50 ms, whereas the bias was non-significant for higher ECG-based SDNN observations. For RMSSD and SDNN, the 95 % LOA was within the a priori 150 % LOA, with 97 % and 100 % of the observed differences lying within the a priori 150 % LOA bounds, respectively. Therefore, the E4 wristband's estimation of RMSSD and SDNN may be considered sufficiently in agreement with the same metrics estimated by the ECG device.

For LF and HF, a tendency to overestimate these metrics by the E4 wristband relative to the ECG device was observed. Namely, in both cases, the entire 95 % CI of the bias was above zero. For LF, the 95 % LOA were within the a priori 150 % LOA for 89 % of the ECG-based LF observations, with acceptable 95 % LOA for observations of ECG-based LF lower than 1979ms^2. Here, 97 % of the observed differences were lying within the a priori 150 % LOA bounds. As such, the E4 wristband estimation of LF was in sufficient agreement with its estimation by the ECG device. However, for HF, the 95 % LOA were within the a priori 150 % LOA for 46 % of the ECG-based HF observations, with acceptable 95 % LOA for observations of ECG-based HF lower than 432ms^2. Here, only 87 % of the observed differences were within the a priori 150 % LOA. This finding illustrated that the E4 wristband's estimation of HF deviated too much from its gold standard estimation by the ECG device.

Consequently, the mean PR, RMSSD, SDNN, and LF were retained for the last step of the three-step hierarchical procedure.

3.1.3. Step 3, the magnitude of difference

The observed effects sizes were negligible for mean PR/HR (Cliff's $d = -0.03$, 95 % CI [−0.28, 0.23]). For the time-domain metrics, the effect sizes were small for RMSSD (Cliff's $d = 0.30$, 95 % CI [0.04, 0.52]) and negligible for SDNN (Cliff's $d = 0.10$, 95 % CI [−0.15, 0.35]). For the frequency-domain metric LF, the effect size was negligible (Cliff's $d = -0.01$, 95 % CI [−0.27, 0.25]). These findings show that the E4 wristband provided comparable estimates of mean PR, RMSSD, SDNN, and LF to their same mean HR and HRV estimation with the ECG device.

3.2. Validity of E4 wristband-based PRV obtained with UST intervals

3.2.1. E4 wristband UST interval PRV versus 5 min E4 wristband-based PRV

3.2.1.1. Step 1, PRV metric selection. The correlation coefficients and

Table 2
Bland-Altman Analysis: Bias, 95 % LOA, and A Priori 150 % LOA.

Metrics	Bias [95% CI]	Lower LOA [95% CI]	Upper LOA [95% CI]	*A priori* 150% LOA
PR/HR*	-0.06 [-0.13, -0.006]	-1.06 [-1.57, -0.99]	0.59 [0.42, 1.08]	-41.10, 41.10
RMSSD*	5.99 [4.56, 7.42]	-2.66 [-5.15, -0.19]	14.65 [12.17, 17.13]	-16.67, 16.67
SDNN*	3.62 - 0.04*GS [2.10, 6.49], [-0.11, -0.004]	Bias - 0.12*GS [0.09, 0.16]	Bias + 0.12*GS [0.09, 0.16]	-21.46, 21.46
LF*	21.69 [0.13, 61.32]	Bias - 0.26*GS [0.18, 0.33]	Bias + 0.26*GS [0.18, 0.33]	-533.21, 533.21
HF	91.60 [37.00, 129.30]	Bias - 0.52*GS [0.38, 0.66]	Bias + 0.52*GS [0.38, 0.66]	**-318.86, 318.86**

Note. PR, mean pulse rate (bpm); HR, mean heart rate (bpm); RMSSD, root mean square of successive differences between normal IBIs (ms); SDNN, standard deviation of normal IBIs (ms); LF, low frequency (ms²), HF, high frequency (ms²); *PR/HR distribution characteristics*, homoscedastic and non-normal which a log transform could not alleviate; *RMSSD distribution characteristics*, homoscedastic and normal; *SDNN distribution characteristics*, heteroscedastic and non-normal which a log transform could not alleviate; *LF distribution characteristics*, heteroscedastic and non-normal which log transformation could not alleviate; *HF distribution characteristics*, heteroscedastic and non-normal which a log transform could not alleviate (see supplementary material E for the assumption handling); Median; 2.5 percentile; 97.5 percentile; β_0, the intercept; β_1, the slope coefficient; antilog, the antilog slope value; GS, the gold standard value; LOA, limits of agreement; 95% CI, 95% confidence intervals; *a priori* 150% LOA, *a priori* 150% limits of agreement; bold typeface indicates when the 95% LOA is outside the *a priori* 150% LOA; * indicates the proxies that were retained for the following step in the three-step hierarchical procedure.

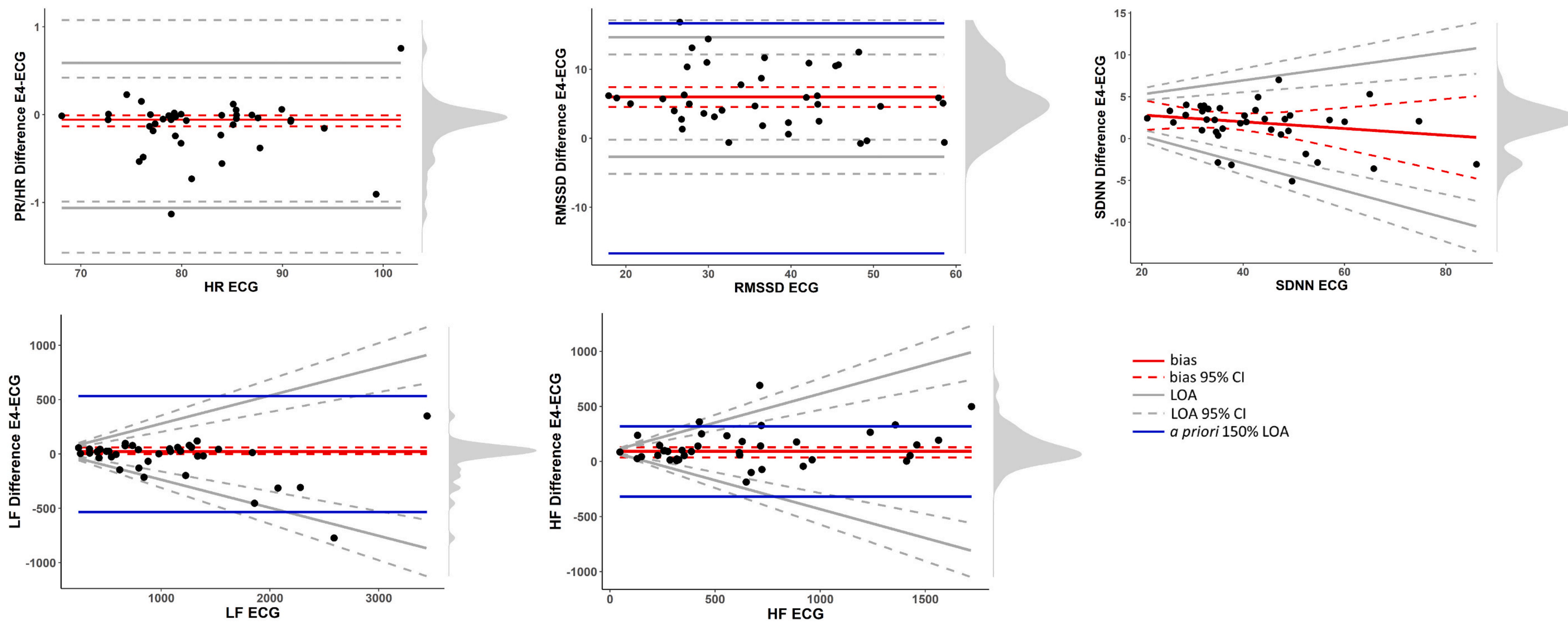

Fig. 2. Bland-Altman plots for mean PR/HR and PRV/HRV metrics.

Note. PR, mean pulse rate (bpm); HR, mean heart rate (bpm); RMSSD, root mean square of successive differences between normal IBIs (ms); SDNN, standard deviation of normal IBIs (ms); LF, low frequency (ms^2), HF, high frequency (ms^2); density distribution of the differences at the right side of each plot; 95 % CI, 95 % confidence interval; LOA, limits of agreement; a priori LOA of 150 % (only displayed if close to 95 % LOA).

Table 3
Correlation coefficients and descriptive statistics of the PR and PRV metrics for the UST and 5 min recordings obtained with the E4 wristband.

Metrics	*r*(*95 % CI*)	*M*(*SD*)
PR		
5 min		82.06(7.13)
10s T1	0.90[0.83, 0.98]	82.22(7.80)
10s T2	0.89[0.82, 0.98]	82.00(8.00)
10s T3	0.93[0.88, 0.98]	81.27(8.51)
Average of 10s	0.96[0.93, 0.99]	81.83(7.67)
30s	0.90[0.82, 0.99]	82.19(7.40)
120 s	0.98[0.96, 1.00]	82.05(6.74)
RMSSD		
5 min		39.33(10.65)
10s T1	0.44[0.15, 0.74]	39.64(15.55)
10s T2	0.70[0.53, 0.85]	36.78(14.95)
10s T3	0.67[0.51, 0.82]	36.27(14.42)
Average of 10s	0.80[0.64, 0.96]	37.56(11.26)
30s	0.79[0.67, 0.90]	37.72(12.23)
120 s	0.93[0.90, 0.97]	39.80(10.94)
SDNN		
5 min		44.50(13.29)
10s T1	0.47[0.20, 0.77]	49.24(20.85)
10s T2	0.61[0.42, 0.80]	38.84(18.22)
10s T3	0.76[0.57, 0.98]	39.93(17.59)
Average of 10s	0.79[0.63, 0.99]	42.67(14.47)
30s	0.78[0.65, 0.94]	42.83(15.20)
120 s	0.93[0.88, 0.99]	45.59(14.44)
HF		
5 min		764.77(518.23)
120 s	0.84[0.75, 0.95]	725.48(511.31)

Note. PR, mean pulse rate (bpm); RMSSD, root mean square of successive differences between normal IBIs (ms); SDNN, standard deviation of normal IBIs (ms); HF, high frequency (ms^2); T, time interval; *r*, Pearson correlation coefficient; 95 % CI = 95 % confidence interval.

descriptive statistics of the mean PR and the PRV metrics of the UST intervals are presented in Table 3. Correlations between UST intervals and 5 min intervals surpassing the cut-off of $r = 0.70$ were found for all UST intervals estimating PR, RMSSD, SDNN, and HF, except for two of the three UST intervals of 10s for RMSSD and SDNN. Hence, all UST intervals estimating PR and PRV surpassing the cut-off were included in the second step of the three-step hierarchical procedure.

3.2.1.2. Step 2, PRV metric validity. The results of the Bland-Altman analysis are depicted in Table 4, and exemplary Bland-Altman plots are shown in Fig. 3 (see supplementary material E for the assumptions handling). The bias included zero for all UST intervals estimating mean PR in its 95 % CI, except for the UST interval of 120 s. Here, proportional bias was detected (i.e., a negative association between the observed differences and the 5 min interval estimation of mean PR; see Fig. 3). Specifically, PR was slightly underestimated by the UST interval of 120 s for observations of 5 min-based PR higher than 96 bpm, whereas the bias was non-significant for lower 5 min-based PR values. For all other UST intervals, no tendency to over/underestimate mean PR was found relative to its same estimation with the 5 min interval. For all UST intervals' estimation of mean PR, the 95 % LOA was inside the a priori 150 % LOA, thereby showing that the UST intervals provided comparable mean PR values to the same value acquired with the 5 min interval. In all cases, 100 % of the data lay within the a priori 150 % LOA.

For all the UST intervals estimating RMSSD, the bias included zero in its 95 % CI. Therefore, for all these UST intervals, no tendency to over/underestimate RMSSD relative to the same estimation with the 5 min interval was found. For the UST interval of 30s, the 95 % LOA was within the a priori 150 % LOA for 85 % of the 5 min-based RMSSD observations, with acceptable 95 % LOA for observations of 5 min-based RMSSD lower than 51 ms. Here, 97 % of the observed differences were lying within the a priori 150 % LOA bounds. For the UST intervals of 120 s, the 95 % LOA were inside the a priori 150 %, with 100 % of the observed differences inside the a priori 150 % LOA boundaries. For the average of the three UST intervals of 10s, the lower bound of the 95 % LOA was borderline outside the lower bound of the a priori 150 % LOA (see Fig. 3), with 97 % of the observed differences within the a priori 150 % LOA boundaries. Therefore, the RMSSD estimated with the average of the three UST intervals of 10s, the UST intervals of 30s and 120 s were in sufficient agreement with its estimation by the 5 min interval. However, for the UST intervals of 10s (i.e., T2), the 95 % LOA was outside the a priori 150 % LOA, with only 90 % of the observed differences within the a priori 150 % LOA boundaries. Therefore, the estimation of RMSSD with this UST interval of 10s deviated too much from the same value acquired with the 5 min interval, indicating insufficient agreement.

For all UST intervals estimating SDNN, the bias included zero in its 95 % CI, except for the UST intervals of 10s. Here, the upper bound of the 95 % CI was below zero, indicating a tendency of this UST interval to underestimate SDNN relative to its estimation with the 5 min interval. For all other UST intervals estimating SDNN, no tendency to over/underestimate it was found relative to its same estimation with the 5 min interval. The 95 % LOA was within the a priori 150 % LOA for the average of the three UST intervals of 10s, the UST interval of 30s, and 120 s. For these UST intervals, 97 %, 97 %, and 100 % of the observed differences were inside the a priori 150 % LOA boundaries, respectively. This result showed that the SDNN estimated by the average of the three UST intervals of 10s, the UST interval of 30s and 120 s was in sufficient agreement with its same estimation by the 5 min interval. For the UST interval of 10s, the 95 % LOA were within the a priori 150 % LOA for 8 % of the 5 min-based SDNN observations, with acceptable 95 % LOA for observations of 5 min-based SDNN lower than 31 ms. Here, only 90 % of the observed differences were within the a priori 150 % LOA boundaries. As such, the SDNN estimated with the UST interval of 10s deviated too much from its estimation with the 5 min interval, thereby showing insufficient agreement.

For the UST interval of 120 s estimating HF, the upper bound of the bias's 95 % CI was below zero, indicating a tendency of this UST interval to underestimate HF relative to its same estimation with the 5 min interval (see Fig. 3). Furthermore, the 95 % LOA were within the a priori 150 % LOA for 8 % of the 5 min-based HF observations, with acceptable 95 % LOA for observations of 5 min-based HF lower than $274ms^2$. Here, only 82 % of the observed differences were within the a priori 150 % LOA boundaries. Therefore, the HF estimated with the UST interval of 120 s deviated too much from its estimation with the 5 min interval, indicating insufficient agreement.

Consequently, all UST intervals estimating mean PR, the average of the three UST intervals of 10s, the UST interval of 30s and 120 s estimating RMSSD and SDNN, were further analyzed in the final step of the three-step hierarchical procedure.

3.2.1.3. Step 3, magnitude of difference. Regarding mean PR, the effect sizes were negligible for all the UST intervals: T1 (Cliff's $d = -0.21$, 95 % CI [−0.27, 0.23]), T2 (Cliff's $d = 0.03$, 95 % CI [−0.22, 0.28]), T3 (Cliff's $d = 0.11$, 95 % CI [−0.15, 0.35]), the average of three 10s intervals (Cliff's $d = 0.03$, 95 % CI [−0.22, 0.27]), 30s interval (Cliff's $d = -0.01$, 95 % CI [−0.26, 0.24]) and 120 s interval (Cliff's $d = -0.01$, 95 % CI [−0.27, 0.24]. Concerning RMSSD, negligible effect sizes were found for the average of the three UST intervals of 10s (Cliff's $d = 0.10$, 95 % CI [−0.15, 0.34]), the UST interval of 30s (Cliff's $d = 0.12$, 95 % CI [−0.13, 0.36], and the UST interval of 120 s (Cliff's $d = -0.04$, 95 % CI [−0.29, 0.22]. Lastly, for the UST intervals estimating SDNN negligible effect sizes were found for the average of the three UST intervals of 10s (Cliff's $d = 0.10$, 95 % CI [−0.16, 0.34]), the UST interval of 30s (Cliff's $d = 0.04$, 95 % CI [−0.22, 0.29]), and the UST interval of 120 s (Cliff's $d = -0.04$, 95 % CI [−0.29, 0.22]). These results show that the UST intervals specified above provide estimates of mean PR, RMSSD and SDNN comparable to the same value estimated with a 5 min interval.

Table 4
UST Recording's Bland-Altman Analysis: Bias and 95 % LOA.

Metrics	Bias [95% CI]	Lower LOA [95% CI]	Upper LOA [95% CI]	*A priori* 150% LOA
PR				
10s T1*	0.17 [-0.93, 1.25]	-6.42 [-8.31, -4.54]	6.75 [4.86, 8.64]	-41.03, 41.03
10s T2*	-0.06 [-1.22, 1.10]	-7.08 [-9.09, -5.07]	6.96 [4.95, 8.97]	-41.03, 41.03
10s T3*	-0.79 [-1.84, 0.26]	-7.15 [-8.98, -5.33]	5.57 [3.75, 7.39]	-41.03, 41.03
Average of 10s*	-0.23 [-0.93, 0.47]	-4.47 [-5.69, -3.26]	4.02 [2.80, 5.23]	-41.03, 41.03
30s*	0.13 [-0.91, 1.18]	-6.19 [-8.00, -4.38]	6.46 [4.65, 8.27]	-41.03, 41.03
120s*	6.22 - 0.08*GS [0.64, 11.80], [-0.14, -0.01]	Bias - 1.96*1.45 [1.18, 1.79]	Bias + 1.96*1.45 [1.18, 1.79]	-41.03, 41.03
RMSSD				
10s T2	-4.58 [-6.77, 0.82]	-25.15 [-38.02, -22.10]	23.96 [23.75, 34.21]	**-19.67, 19.67**
Average of 10s*	-1.33 [-4.79, 1.15]	-21.24 [-34.77, -16.89]	15.26 [12.06, 24.91]	**-19.67,** 19.67
30s*	-1.62 [-4.02, 0.56]	Bias - 0.36*GS [0.26, 0.45]	Bias + 0.36*GS [0.26, 0.45]	-19.67, 19.67
120s*	0.47 [-0.85, 1.79]	-7.49 [-9.77, -5.21]	8.43 [6.15, 10.71]	-19.67, 19.67
SDNN				
10s T3	-4.57 [-8.31, -0.84]	Bias – 2.46*(1.92 + 0.17*GS) [-5.20, 9.03], [0.01, 0.32]	Bias + 2.46*(1.92 + 0.17*GS) [-5.20, 9.03], [0.01, 0.32]	**-22.25, 22.25**
Average of 10s*	-1.83 [-4.76, 1.10]	-19.55 [-24.62, -14.47]	15.88 [10.81, 20.96]	-22.25, 22.25
30s*	-1.67 [-4.77, 1.42]	-20.38 [-25.74, -15.02]	17.04 [11.68, 22.39]	-22.25, 22.25
120s*	1.09 [-0.66, 2.84]	-9.50 [-12.54, -6.47]	11.68 [8.65, 14.71]	-22.25, 22.25
HF				
120s	-39.29 [-132.9, -54.30]	Bias – 2.46*(98.10 + 0.15*GS) [7.07, 189.10], [0.05, 0.25]	Bias + 2.46*(98.10 + 0.15*GS) [7.07, 189.10], [0.05, 0.25]	**-382.39, 382.39**

Note. PR, mean pulse rate (bpm); RMSSD, root mean square of successive differences between normal IBIs (ms); SDNN, standard deviation of normal IBIs (ms); HF, high frequency (ms²); T = time interval; *PR UST intervals distribution characteristics*, all were homoscedastic and normally distributed with only the UST interval of 120s showing proportional bias; *RMSSD UST distribution characteristics*, all were homoscedastic with T2, the average of the three UST interval of 10s, and the UST interval of 30s displaying non-normality for which log transformation only alleviated non-normality for the UST interval of 30s; *SDNN UST intervals distribution characteristics*, all were homoscedastic and normally distributed except for the UST interval of 10s which was heteroscedastic; *HF UST interval distribution characteristics*, data was heteroscedastic and normally distributed (see supplementary material E for the assumption handling); Median; 2.5 percentile; 97.5 percentile; β_0, the intercept; β_1, the slope coefficient; Bias *SD*, the *SD* of the residuals of the proportional bias model; antilog, the antilog slope value; GS, the gold standard value; LOA = 95% limits of agreement; 95% CI = 95% confidence interval; *A priori* 150% LOA, *a priori* 150% limits of agreement; Bold typeface = the 95% LOA is outside the *a priori* 150% LOA; * indicates the proxies that were retained for the following step in the three-step hierarchical procedure.

3.2.2. E4 wristband UST interval PRV versus 5 min ECG-based HRV

3.2.2.1. Step 1, PRV metric selection. For all UST intervals, strong correlations ($r > 0.70$) were observed between their estimation of PR and PRV and the 5 min ECG-based estimation of HR and HRV, except for the three UST intervals of 10s for RMSSD and two of the three UST intervals of 10s for SDNN.

3.2.2.2. Step 2, PRV metric validity. Considering only those UST intervals that survived step 1, all UST intervals' estimations of PR were in sufficient agreement with the 5 min ECG-based estimation of HR. Concerning RMSSD, only its estimation with the UST interval of 30s was in sufficient agreement with the same estimation with the 5 min ECG recording. For the average of the three UST intervals of 10s and the UST interval of 120 s, only 92 % of the observed differences were within the a priori 150 % LOA. Regarding SDNN, only its estimation with the average of the three UST intervals of 10s and the UST interval of 30s and 120 s were in sufficient agreement with their same estimation with the 5 min ECG recording. For the SDNN estimation with the UST interval of 10s, only 90 % of the observed differences were within the a priori 150 % LOA. Concerning HF, its estimation with the UST interval of 120 s agreed insufficiently with the same estimation with the 5 min ECG recording.

3.2.2.3. Step 3, magnitude of difference. Considering only those UST intervals that survived step 2, negligible effect sizes were found between the UST intervals estimation of PR and SDNN and the 5 min ECG-based estimation of HR and SDNN. A small effect size was found for the UST interval of 30s estimating RMSSD (see supplementary material D for the detailed statistical results).

4. Discussion

The current study aimed to validate the mean PR and PRV metrics obtained with the E4 wristband as approximations of mean HR and HRV metrics by comparing them to the mean HR and HRV metrics acquired with a gold standard ECG device. Moreover, we assessed the time scales at which the E4 wristband can validly derive PR and PRV by comparing its UST interval recordings of mean PR and PRV metrics to that of a gold standard 5 min interval recording. To achieve these two aims, participants' IBIs were simultaneously recorded with an E4 wristband and an

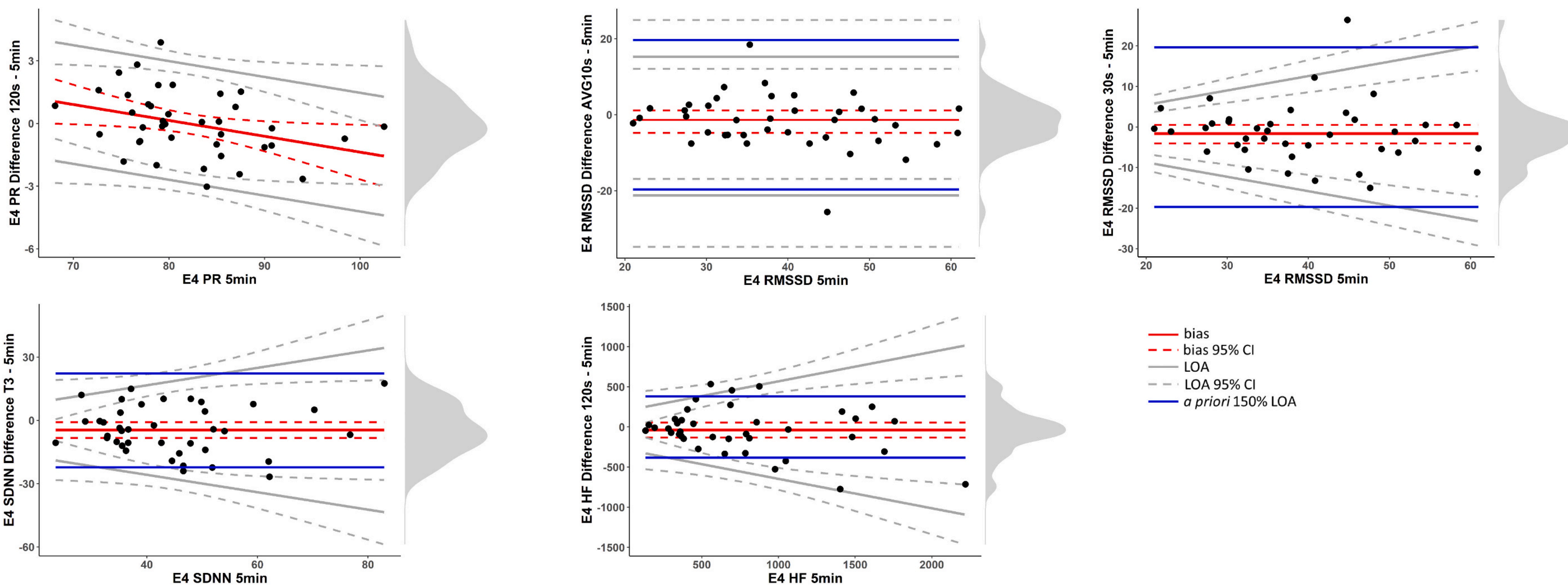

Fig. 3. Examples of Bland-Altman plots comparing mean PR and PRV metrics obtained with UST intervals by the E4 wristband versus their 5 min recording with the E4 wristband.
Note. PR, mean pulse rate (bpm); RMSSD, root mean square of successive differences between normal IBIs (ms); SDNN, standard deviation of normal IBIs (ms); HF, high frequency (ms^2); AVG10s, average of the three 10s intervals; T3, third UST interval of 10s; density distribution of the differences at right side plot; 95 % LOA, 95 % limit of agreement; 95 % CI, 95 % confidence intervals; the a priori LOA of 150 % only displayed if close to 95 % LOA.

ECG device during a 5 min seated-rest condition.

4.1. Validity of E4 wristband-based PRV

With regards to the E4 wristbands' PRV metrics validity as approximations of HRV metrics, our results are largely consistent with previous studies. Similar to Menghini et al. (2019), we found that, in a seated condition, the E4 wristband's estimation of mean PR, RMSSD, SDNN, and LF were comparable to their same mean HR and HRV estimation with the gold standard ECG device (see also van Lier et al., 2020). However, we found that HF was invalidly estimated by the E4 wristband. This finding is inconsistent with some results (e.g., Menghini et al., 2019; Schuurmans et al., 2020) but consistent with others (e.g., Ollander et al., 2016).

In line with Menghini et al. (2019), we found that the E4 wristband tended to overestimate RMSSD and LF relative to their estimation with the ECG device. This is valuable information for researchers as it allows researchers to potentially apply a calibration index (i.e., subtracting or adding a value) to the E4 wristband's estimation of PRV so that it approaches the true HRV value more precisely. However, the nature of this calibration index depends on whether the bias is proportional; as such, assessing proportional bias is needed.[2] Furthermore, it might be an interesting avenue for further inquiry to assess whether the bias (i.e., systematic error) of the E4 wristband is due to noise or some important underlying physiological parameter inherent to BVP. Analyzing the pulse transit time and pulse wave velocity may offer an opportunity to assess how, for instance, vasodilation/constriction, arterial stiffness, and arterial compliance (i.e., physiological factors influencing blood flow and pressure) are related to this systematic error (e.g., Mejía-Mejía et al., 2021; Mol et al., 2020).

4.2. Validity of E4 wristband-based PRV obtained with UST intervals

The current study is, to our knowledge, the first to assess the fine-grained character of the time scales at which the E4 wristband can validly derive PR and PRV metrics. Our results corroborate many other studies (e.g., Baek et al., 2015; Munoz et al., 2015; Shaffer et al., 2016) in showing that mean PR/HR can be validly assessed at a large range of UST intervals (e.g., 10s). This indicates that the E4 wristband is, in that regard, similar to other recording devices. However, for the time-domain PRV metrics RMSSD and SDNN, the PRV recordings with the three shortest 10s UST intervals were unstable, whereas taking the average of them or using longer UST intervals (i.e., 30s and 120 s) significantly increased stability. This result is not in line with Munoz et al. (2015), who demonstrated that RMSSD could be validly derived with 10s UST intervals (see also Nussinovitch et al., 2011). However, their participants were in a supine-rest condition as opposed to the seated-rest condition of the current study, perhaps making their recording more stable due to fewer motion artifacts. Other studies using a seated-rest condition also observed diminished stability of, for instance, RMSSD obtained with UST recording intervals of 10s (e.g., Baek et al., 2015; Salahuddin et al., 2007; Shaffer et al., 2016). Lu et al. (2008) compared PRV to HRV during supine and seated rest conditions and showed a diminishing accuracy during seated as compared to supine rest. The instability in the data might increase at such UST intervals of 10s because fluctuations in the PPI time series might occur precisely in the 10s recording, distorting the estimation of the time-domain PRV metrics RMSSD and SDNN. In longer recording intervals (e.g., 30s, 120 s, 5 min) or when averaging over multiple UST intervals, these irregular fluctuations in the PPI time series more likely are leveled out. Thus, averaging over multiple short intervals or increasing the UST recording length to 30s and 120 s improved the accuracy of estimating RMSSD and SDNN with the E4 wristband. Especially for SDNN, this is not that surprising as it is a measure of the variability of the IBIs. This variability is expected to increase with the recording length (i.e., the more data, the more variability; Task Force, 1996). Our RMSSD and SDNN observations mostly mimic the results of other studies (e.g., Munoz et al., 2015; Shaffer et al., 2016). Although for the estimation with the UST interval of 30s, some found them to deviate too much from their 5 min estimation (e.g., Shaffer et al., 2016), whereas others did find that these were reliable proxies (e.g., Baek et al., 2015; Munoz et al., 2015; Salahuddin et al., 2007). In general, for those UST intervals deemed valid in estimating mean PR and PRV metrics, as specified above, there was no tendency to over/underestimate mean PR or the PRV metrics relative to their same estimation with a 5 min interval.

The HF obtained with the UST interval of 120 s agreed insufficiently with its value acquired with 5 min interval recording. This result is consistent with the finding of Shaffer et al. (2016), who found that HF could only be validly estimated with a UST interval of 180 s. Even though others did find HF to be validly estimated with UST intervals of 20s and 40s (Baek et al., 2015; Salahuddin et al., 2007), we consider this implausible as a proper estimation of power in the HF band requires at least 70s (Task Force, 1996). This 70s minimum only ensures sufficient time to decently perform a power calculation in the HF band. This does not automatically imply that this is sufficiently long to estimate HF validly. As the current study examined BVP-based estimation of HF, which is less stable than ECG-based HF, even doubling this bare minimum 70s proved not to be sufficient. So based on our results, we recommend that longer BVP recordings are required for HF.

It is noteworthy in our second UST analysis, where we compared the estimation of the UST intervals of PR and PRV to the 5 min ECG-based HR and HRV (see supplementary material D), that for RMSSD only the estimation with the UST interval of 30s was found valid. Therefore, the BVP-based estimation of RMSSD with UST intervals seems a less stable surrogate for ECG-based RMSSD with a 5 min interval. On closer inspection of the Bland-Altman plots, we observed that, for the average of the three UST intervals of 10s and the UST interval of 120 s, a large part (i.e., 92 %) of these UST intervals' RMSSD estimation was acceptable, and only three observations mainly caused this invalidity with an E4 wristband-based inaccuracy of ±20 ms. The discrepancy between the two UST analyses might be related to the BVP-related systematic error. As all measurements are BVP-based in the first analysis, this systematic error is accounted for. However, this is not the case in this second analysis. Here, systematic and random errors are combined, which might lead to differing results. The findings of the second UST analysis illustrate that BVP-based PRV mostly approximates ECG-based HRV. However, there are still marked differences between the two, likely due to different physiological constraints, as BVP-based PRV is biomechanical whereas ECG-based HRV is bio-electrical (see Yuda et al., 2020, for argumentation).

4.3. Statistical procedures

One source that leads to inconsistencies in research results and difficult comparability between studies is the statistical procedure employed to assess agreement between the proxy and the gold standard. Although commendable attempts have been made to establish valid statistical protocols for the assessment of agreement (e.g., Bland and Altman, 1999; Menghini et al., 2019; Menghini et al., 2021; Pecchia et al., 2018; van Lier et al., 2020), they still rely on some procedures that have been criticized such as correlational analyses and Cohen's *d* tests based on paired samples *t*-tests (e.g., Pecchia et al., 2018). For example, correlational analyses highlight the strength of an association but no consistent up/downward shift of the values of one of the variables relative to the other variable. We followed the proposed protocols to

[2] For example, the E4 wristband's tendency to overestimate SDNN was negatively associated with the size of the ECG-based SDNN, implying that the calibration index shifts depending on ECG-based SDNN. Therefore, one would need the coefficients of the proportional bias model (i.e., calibration index = $\beta_0 + \beta_1$*ECG-based value) to calculate the required calibration index.

illustrate consistency in agreement patterns across several statistical procedures. We argue the Bland-Altman analysis should be the main source of information to identify agreement. However, not all studies rely on Bland-Altman analysis (e.g., Ollander et al., 2016) or do so without using a priori LOAs and/or indicating how Bland-Altman assumptions were dealt with (e.g., Schuurmans et al., 2020; Kiran Kumar et al., 2021). As these assumptions strongly influence the calculation and representation of the 95 % LOA, it is difficult to interpret the results of these studies unambiguously.

Moreover, the a priori LOA calculation is not consistent across the literature, and its proposed cut-off differs and is relatively arbitrary (i.e., 110 % or 150 %). Menghini et al. (2019) take ±50 % of the mean of the gold standard, whereas van Lier et al. (2020) take 10 % of the range of biologically plausible HRV values, which can vary between age cohorts (e.g., 20–29 years or 50–60 years; Umetani et al., 1998). As highlighted in the Methods section, we argue, similarly to Giavarina (2015), that this choice should depend on the goal for which the E4 wristband or UST interval estimations of PRV are used. For instance, for medical diagnostic purposes, the accuracy of PRV estimation as an approximation of HRV should be high (e.g., a priori LOA of 110 %), whereas for lab-based research, it can be less strict (e.g., a priori LOA of 150 %).

4.4. Recommendations

E4 wristband validation studies examining PRV as approximations of HRV metrics and time scale at which PRV can validly be derived are limited or non-existing and use divergent procedures. Therefore, more procedurally consistent E4 wristband validation studies are needed.

Our study concludes that using the E4 wristband as a research-grade device to track PR/HR and estimate PRV/HRV in seated conditions in a lab-based context seems valuable under certain conditions. The E4 wristband provided comparable measures to the gold standard ECG device concerning mean HR, RMSSD, SDNN, and LF. This observation indicates that, in a lab setting where participants are seated and there is limited movement of the hand wearing the wristband, the E4 wristband can be a valid substitute for an ECG device with a 5 min recording length. However, several observations need to be made.

First, the BVP signal was clearly less stable than the ECG signal (see supplementary material C), which led to a substantial exclusion of participants. Hence, the potential advantage of the E4 wristband to facilitate meeting sample size requirements more easily than with an ECG device might be canceled out by the sample size compensation needed to accommodate the instability of the BVP signal. However, as the E4 wristband is considerably cheaper, one can test multiple participants simultaneously. So even though one might need to exclude 32 % of the data, it still may be more labor efficient to work with the E4 wristband than an ECG device. On the other hand, ethically, we must question whether designs, where 32 % of the participants will have to be removed, are justified, especially in demanding task situations. We did not explicitly tell participants to keep the hand wearing the E4 wristband completely at rest during the baseline recording. It was only indicated to remain calm and control their emotions without excess movements (e.g., arm stretching). This perhaps caused an unstable BVP signal in some cases. On the other hand, as participants are likely to have made some minimal movement during the 5 min rest, it shows that the E4 wristband's PR and PRV estimation as approximations of HR and HRV estimation was relatively valid under such minimal-movement conditions. We also asked participants to put on the E4 wristband themselves. Although we visually checked if it was put on correctly, it may have been that this was, on some occasions, not attached firmly enough, causing an unstable BVP signal. To enhance the quality of the signal, one could (1) ask participants to keep the hand wearing the E4 wristband completely at rest and (2) let the experimenter attach the E4 wristband watch to ensure that it is attached correctly and firmly. Furthermore, increasing the recording length might also remedy this. Longer BVP recording intervals lead to more stable BVP signal intervals that can be used to validly estimate the mean HR and HRV metrics (see, for example, So et al., 2021). Lastly, it is important not to rely solely on BVP preprocessing algorithms (e.g., Kubios, Tarvainen et al., 2020) to correct artifacts and noise segments in the BVP signal. During preprocessing, we noticed that these algorithms sometimes missed uninterpretable noise segments. Therefore, we recommend a visual check of the BVP signal for its morphology and stability to ensure proper cleaning/preprocessing of the BVP signal. Moreover, it is advisable to include a rest period between trials of at least 2 min (the current study used a 5 min acclimatization period) to enhance the stationarity of the BVP data and, thus, its quality.

Second, we assessed the validity of the E4 wristband only in a condition involving minimal movement. But even then, we could not completely avoid the instability of the BVP signal. This observation possibly constrains the possibilities of using the E4 wristband in real-life situations. If movement is involved (e.g., walking or running), the E4 wristband is probably less likely to provide valid estimates. Still, other real-life situations might fulfill the E4 wristband's preconditions, such as studying HRV in classrooms or workplaces, at the bedside of patients, or at the home of older adults, where participants are seated or lying down and moving minimally. Furthermore, we only tested participants in a seated resting state. Thus, based on our results, we cannot assert how valid the E4 wristband is when, for instance, performing a computer task. However, future research-based studies can now implement the E4 wristband to assess its validity, for example, during task completion, under stress or when facing cognitive load. The recommendations we provide here can help to promote a valid E4 wristband data acquisition in a lab-based context. We also note that the current validation procedure was part of a broader creative problem solving experiment, of which the experimental setup might have had an undue influence on the validation procedure. However, Shcheslavskaya et al. (2010) have shown that the heart's activity returns to baseline within 6 min after a cognitive challenge (see also Panaite et al., 2015). Although participants had to read instructions and perform six practice trials before the ECG and BVP recording, there was a 5 min acclimatization period before the actual recording started, which should have been sufficient to eliminate any cardiovascular reactivity caused by the instructions and/or practice trials.

Third, our sample consisted almost exclusively of female undergraduates. Although this observation limits the scope in which our results can be interpreted, it is noteworthy that our results were similar to some other studies with a more biological sex-balanced sample (e.g., Menghini et al., 2019; van Lier et al., 2020) and having a broader age range (e.g., Menghini et al., 2019). Future studies should also consider recording additional demographic information, as it has been argued that, for instance, skin tone could impact PPG-based PRV recording (Fallow et al., 2013), although this has not been consistently reported (e.g., Bent et al., 2020).

Fourth, concerning the UST intervals of the E4 wristband, our results showed that the E4 wristband's time scales to derive PRV were not extremely fine-grained. Deriving PRV metrics based on 10s UST intervals did not produce valid estimates. However, the E4 wristband did prove its usefulness for lab experiments across all UST intervals for mean PR and when averaging over multiple short-duration intervals (e.g., 10s) or when using a recording length of at least 30s for RMSSD and SDNN. We must note that, solely for RMSSD, the UST intervals' comparison with 5 min ECG only revealed a valid RMSSD with the UST interval of 30s. Although most RMSSD observations with the UST intervals were valid, a few deviated strongly from the 5 min ECG-based RMSSD. Therefore, caution is warranted when using E4 wristband-based RMSSD with UST intervals as a proxy of ECG-based RMSSD. Assessing HF presumably would require a longer recording length than 120 s (see, for example, Shaffer et al., 2016). These observations seem to place certain time constraints on using the E4 wristband in a lab context. For instance, studies aiming to assess PRV on a trial-by-trial basis with short trial durations (e.g., Shen et al., 2017) might not produce valid PRV estimates with the E4 wristband. However, the E4 wristband does lend itself to recording time-domain PRV metrics, for example, on a trial-by-trial basis or across a block of trials if the trial or block lasts at least 30s or if an average can be taken across multiple shorter trials. For the frequency domain PRV metrics, a trial-by-trial analysis seems difficult as HF was found to be invalidly derived with the E4 wristband with a UST interval of 120 s.

To summarize, the E4 wristband is a piece of promising research equipment in seated lab conditions. Our results showed that mean HR, RMSSD, SDNN, and LF were validly estimated by the E4 wristband. Contrary to some studies, we could not corroborate the valid estimation of HF. Furthermore, we showed that the E4 wristband's mean PR was a valid proxy of the 5 min gold standard recording across all UST intervals. RMSSD and SDNN could be validly estimated with the E4 wristband with an average over multiple UST intervals of 10s, a UST interval of 30s or 120 s, but not with 10s UST intervals. However, for RMSSD, when compared to a 5 min ECG recording, only the UST interval of 30s remained valid. For HF, longer recording times seem to be required. Based on these results, we have formulated several recommendations for the use of the E4 wristband in laboratory research contexts.

Funding

This work was supported by the “Fonds de la Recherche Scientifique” [grant number 34736358, 2019] and the Research Foundation Flanders [grant number G096919N].

Data availability

Data will be made available on the Open Science Framework

Acknowledgments

We thank Febe Demeyer, Christo Bratanov, Yujing Liang, and Amar Music for their assistance with the data collection and their critical thought and the Fonds de la Recherche Scientifique and the Research Foundation Flanders (FWO) for providing the opportunity to conduct this research under a research fellow grant.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.ijpsycho.2022.10.003.

References

Ajayi, T.A., Salongo, L., Zang, Y., Wineinger, N., Steinhubl, S., 2021. Mobile health-collected biophysical markers in children with serious illness-related pain. J. Palliat. Med. 24 (4), 580–588. https://doi.org/10.1089/jpm.2020.0234.

Alqaraawi, A., Alwosheel, A., Alasaad, A., 2016. Towards efficient heart rate variability estimation in artifact-induced photoplethysmography signals. In: 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE). IEEE, pp. 1–6. https://doi.org/10.1109/CCECE.2016.7726853.

Altman, D.G., Bland, J.M., 1983. Measurement in medicine: the analysis of method comparison studies. The Statistician 32 (3), 307–317. https://doi.org/10.2307/2987937.

Association for the Advancement of Medical Instrumentation, 2002. Cardiac monitors, heart rate meters, and alarms. In: American National Standard (ANSI/AAMI EC13: 2002) Arlington, VA, pp. 1–87.

Baek, H.J., Cho, C.H., Cho, J., Woo, J.M., 2015. Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability. Telemed. e-Health 21 (5), 404–414. https://doi.org/10.1089/tmj.2014.0104.

Bent, B., Goldstein, B.A., Kibbe, W.A., Dunn, J.P., 2020. Investigating sources of inaccuracy in wearable optical heart rate sensors. Npj Digit. Med. 3 (1), 1–9. https://doi.org/10.1038/s41746-020-0226-6.

Berntson, G.G., Thomas Bigger Jr., J., Eckberg Jr., D.L., Grossman Jr., P., Kaufmann Jr., P.G., Malik Jr., M., Van Der Molen, M.W., 1997. Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology 34 (6), 623–648. https://doi.org/10.1111/j.1469-8986.1997.tb02140.x.

Bland, J.M., Altman, D.G., 1995. Comparing methods of measurement: why plotting difference against standard method is misleading. The Lancet 346 (8982), 1085–1087. https://doi.org/10.1016/S0140-6736(95)91748-9.

Bland, J.M., Altman, D.G., 1999. Measuring agreement in method comparison studies. Stat. Methods Med. Res. 8 (2), 135–160. https://doi.org/10.1177/096228029900800204.

Campbell, J.I., Thompson, V.A., 2012. MorePower 6.0 for ANOVA with relational confidence intervals and bayesian analysis. Behav. Res. Methods 44 (4), 1255–1265. https://doi.org/10.3758/s13428-012-0186-0.

Castaneda, D., Esparza, A., Ghamari, M., Soltanpur, C., Nazeran, H., 2018. A review on wearable photoplethysmography sensors and their potential future applications in health care. Int. J. Biosens. Bioelectron. 4 (4), 195. https://doi.org/10.15406/ijbsbe.2018.04.00125.

Charlot, K., Cornolo, J., Brugniaux, J.V., Richalet, J.P., Pichon, A., 2009. Interchangeability between heart rate and photoplethysmography variabilities during sympathetic stimulations. Physiol. Meas. 30 (12), 1357. https://doi.org/10.1088/0967-3334/30/12/005.

Choi, K.H., Kim, J., Kwon, O.S., Kim, M.J., Ryu, Y.H., Park, J.E., 2017. Is heart rate variability (HRV) an adequate tool for evaluating human emotions?–A focus on the use of the international affective picture system (IAPS). Psychiatry Res. 251, 192–196. https://doi.org/10.1016/j.psychres.2017.02.025.

Euser, A.M., Dekker, F.W., le Cessie, S., 2008. A practical approach to bland-altman plots and variation coefficients for log transformed variables. J. Clin. Epidemiol. 61 (10), 978–982. https://doi.org/10.1016/j.jclinepi.2007.11.003.

Fallow, B.A., Tarumi, T., Tanaka, H., 2013. Influence of skin type and wavelength on light wave reflectance. J. Clin. Monit. Comput. 27 (3), 313–317. https://doi.org/10.1007/s10877-013-9436-7.

Giavarina, D., 2015. Understanding bland altman analysis. Biochem. Med. 25 (2), 141–151. https://doi.org/10.11613/BM.2015.015.

Graham, S.A., Jeste, D.V., Lee, E.E., Wu, T.C., Tu, X., Kim, H.C., Depp, C.A., 2019. Associations between heart rate variability measured with a wrist-worn sensor and older adults' physical function: observational study. JMIR mHealth uHealth 7 (10), e13757. https://doi.org/10.2196/13757.

Heathers, J.A., 2014. Everything hertz: methodological issues in short-term frequency-domain HRV. Front. Physiol. 5, 177. https://doi.org/10.3389/fphys.2014.00177.

Ishaque, S., Khan, N., Krishnan, S., 2021. Trends in heart-rate variability signal analysis. Front. Digit. Health 3, 13. https://doi.org/10.3389/fdgth.2021.639444.

Kiran Kumar, C., Manaswini, M., Maruty, K.N., Siva Kumar, A.V., Mahesh Kumar, K., 2021. Association of heart rate variability measured by RR interval from ECG and pulse to pulse interval from photoplethysmography. Clin. Epidemiol. Glob. Health 10 (100698), 2213–3984. https://doi.org/10.1016/j.cegh.2021.100698.

Krouwer, J.S., 2008. Why bland-altman plots should use X, not (Y + X)/2 when X is a reference method. Stat. Med. 27 (5), 778–780. https://doi.org/10.1002/sim.3086.

Kuipers, M., Richter, M., Scheepers, D., Immink, M.A., Sjak-Shie, E., van Steenbergen, H., 2017. How effortful is cognitive control? Insights from a novel method measuring single-trial evoked beta-adrenergic cardiac reactivity. Int. J. Psychophysiol. 119, 87–92. https://doi.org/10.1016/j.ijpsycho.2016.10.007.

Kumral, D., Schaare, H.L., Beyer, F., Reinelt, J., Uhlig, M., Liem, F., Gaebler, M., 2019. The age-dependent relationship between resting heart rate variability and function brain connectivity. NeuroImage 185, 521–533. https://doi.org/10.1016/j.neuroimage.2018.10.027.

Laborde, S., Mosley, E., Thayer, J.F., 2017. Heart rate variability and cardiac vagal tone in psychophysiological research–recommendations for experiment planning, data analysis, and data reporting. Front. Psychol. 8, 213. https://doi.org/10.3389/fpsyg.2017.00213.

Lackner, H.K., Weiss, E.M., Schulter, G., Hinghofer-Szalkay, H., Samson, A.C., Papousek, I., 2013. I got it! Transient cardiovascular response to the perception of humor. Biol. Psychol. 93 (1), 33–40. https://doi.org/10.1016/j.biopsycho.2013.01.014.

Lipponen, J.A., Tarvainen, M.P., 2019. A robust algorithm for heart rate variability time series artefact correction using novel beat classification. J. Med. Eng. Technol. 43 (3), 173–181. https://doi.org/10.1080/03091902.2019.1640306.

Lu, G., Yang, F., 2009. Limitations of oximetry to measure heart rate variability measures. Cardiovasc. Eng. 9 (3), 119–125. https://doi.org/10.1007/s10558-009-9082-3.

Lu, S., Zhao, H., Ju, K., Shin, K., Lee, M., Shelley, K., Chon, K.H., 2008. Can photoplethysmography variability serve as an alternative approach to obtain heart rate variability information? J. Clin. Monit. Comput. 22 (1), 23–29. https://doi.org/10.1007/s10877-007-9103-y.

McCarthy, C., Pradhan, N., Redpath, C., Adler, A., 2016. Validation of the Empatica E4 wristband. In: 2016 IEEE EMBS International Student Conference (ISC). IEEE, pp. 1–4. https://doi.org/10.1109/EMBSISC.2016.7508621.

Mejía-Mejía, E., May, J.M., Torres, R., Kyriacou, P.A., 2020. Pulse rate variability in cardiovascular health: a review on its applications and relationship with heart rate variability. Physiol. Meas. 41 (7), 07TR01. https://doi.org/10.1088/1361-6579/ab998c.

Mejía-Mejía, E., May, J.M., Elgendi, M., Kyriacou, P.A., 2021. Differential effects of the blood pressure state on pulse rate variability and heart rate variability in critically ill patients. NPJ Digit. Med. 4 (1), 1–11. https://doi.org/10.1038/s41746-021-00447-y.

Menghini, L., Gianfranchi, E., Cellini, N., Patron, E., Tagliabue, M., Sarlo, M., 2019. Stressing the accuracy: wrist-worn wearable sensor validation over different conditions. Psychophysiology 56 (11), e13441. https://doi.org/10.1111/psyp.13441.

Menghini, L., Cellini, N., Goldstone, A., Baker, F.C., de Zambotti, M., 2021. Sleep-tracking technology: step-by-step guidelines and open-source code. Sleep 44 (2), 1–12. https://doi.org/10.1093/sleep/zsaa170.

Milstein, N., Gordon, I., 2020. Validating measures of electrodermal activity and heart rate variability derived from the empatica E4 utilized in research settings that

involve interactive dyadic states. Front. Behav. Neurosci. 14, 148. https://doi.org/10.3389/fnbeh.2020.00148.
Mishra, T., Wang, M., Metwally, A.A., Bogu, G.K., Brooks, A.W., Bahmani, A., Snyder, M. P., 2020. Pre-symptomatic detection of COVID-19 from smartwatch data. Nat. Biomed. Eng. 4 (12), 1208–1220. https://doi.org/10.1038/s41551-020-00640-6.
Mol, A., Slangen, L.R.N., Trappenburg, M.C., Reijnierse, E.M., van Wezel, R.J.A., Meskers, C.G.M., Maier, A.B., 2020. Blood pressure drop rate after standing up is associated with frailty and number of falls in geriatric outpatients. J. Am. Heart Assoc. 9 (7), e014688 https://doi.org/10.1161/JAHA.119.014688.
Munoz, M.L., van Roon, A., Riese, H., Thio, C., Oostenbroek, E., Westrik, I., Snieder, H., 2015. Validity of (ultra-) short recordings for heart rate variability measurements. PloS One 10 (9), e0138921. https://doi.org/10.1371/journal.pone.0138921.
Nuske, H.J., Goodwin, M.S., Kushleyeva, Y., Forsyth, D., Pennington, J.W., Masino, A.J., Herrington, J.D., 2021. Evaluating commercially available wireless cardiovascular monitors for measuring and transmitting real-time physiological responses in children with autism. Autism Res. 15 (1), 117–130. https://doi.org/10.1002/aur.2633.
Nussinovitch, U., Elishkevitz, K.P., Katz, K., Nussinovitch, M., Segev, S., Volovitz, B., Nussinovitch, N., 2011. Reliability of ultra-short ECG indices for heart rate variability. Ann. Noninvasive Electrocardiol. 16 (2), 117–122. https://doi.org/10.1111/j.1542-474X.2011.00417.x.
Ollander, S., Godin, C., Campagne, A., Charbonnier, S., 2016. A comparison of wearable and stationary sensors for stress detection. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, pp. 004362–004366. https://doi.org/10.1109/SMC.2016.7844917.
Ottaviani, C., Zingaretti, P., Petta, A.M., Antonucci, G., Thayer, J.F., Spitoni, G.F., 2018. Resting heart rate variability predicts inhibitory control above and beyond impulsivity. J. Psychophysiol. 33 https://doi.org/10.1027/0269-8803/a000222.
Panaite, V., Salomon, K., Jin, A., Rottenberg, J., 2015. Cardiovascular recovery from psychological and physiological challenge and risk for adverse cardiovascular outcomes and all-cause mortality. Psychosom. Med. 77 (3), 215–226. https://doi.org/10.1097/PSY.0000000000000171.
Pecchia, L., Castaldo, R., Montesinos, L., Melillo, P., 2018. Are ultra-short heart rate variability features good surrogates of short-term ones? State-of-the-art review and recommendations. Healthcare Technol. Lett. 5 (3), 94–100. https://doi.org/10.1049/htl.2017.0090.
Pichon, A., Roulaud, M., Antoine-Jonville, S., de Bisschop, C., Denjean, A., 2006. Spectral analysis of heart rate variability: interchangeability between autoregressive analysis and fast fourier transform. J. Electrocardiol. 39 (1), 31–37. https://doi.org/10.1016/j.jelectrocard.2005.08.001.
Pulopulos, M.M., Vanderhasselt, M.A., De Raedt, R., 2018. Association between changes in heart rate variability during the anticipation of a stressful situation and the stress-induced cortisol response. Psychoneuroendocrinology 94, 63–71. https://doi.org/10.1016/j.psyneuen.2018.05.004.
Quintana, D.S., Heathers, J.A.J., 2014. Considerations in the assessment of heart rate variability in biobehavioral research. Front. Psychol. 5 (JUL), 1–10. https://doi.org/10.3389/fpsyg.2014.00805.
R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. R-project.org.
Rahman, J.S., Gedeon, T., Caldwell, S., Jones, R., Jin, Z., 2021. Towards effective music therapy for mental health care using machine learning tools: human affective reasoning and music genres. J. Artif. Intell. Soft Comput. Res. 11, 5–20. https://doi.org/10.2478/jaiscr-2021-0001.
Romano, J., Kromrey, J.D., Coraggio, J., Skowronek, J., Devine, L., 2006. Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen'sd indices the most appropriate choices. In: Annual Meeting of the Southern Association for Institutional Research, pp. 1–51.
Ryan, W., Conigrave, J.H., Basarkod, G., Ciarrochi, J., Sahdra, B.K., 2019. When Is It Good to Use Wristband Devices to Measure HRV?: Introducing a New Method for Evaluating the Quality of Data From Photophlethysmography-based HRV Devices. https://doi.org/10.31234/osf.io/t3gdz.
Salahuddin, L., Cho, J., Jeong, M.G., Kim, D., 2007. Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp. 4656–4659. https://doi.org/10.1109/IEMBS.2007.4353378.
Schuurmans, A.A., de Looff, P., Nijhof, K.S., Rosada, C., Scholte, R.H., Popma, A., Otten, R., 2020. Validity of the empatica E4 wristband to measure heart rate variability (HRV) parameters: a comparison to electrocardiography (ECG). J. Med. Syst. 44 (11), 1–11. https://doi.org/10.1007/s10916-020-01648-w.
Shaffer, F., Ginsberg, J.P., 2017. An overview of heart rate variability metrics and norms. Front. Public Health 5, 258. https://doi.org/10.3389/fpubh.2017.00258.
Shaffer, F., McCraty, R., Zerr, C.L., 2014. A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. Front. Psychol. 5, 1040. https://doi.org/10.3389/fpsyg.2014.01040.
Shaffer, F., Shearman, S., Meehan, Z.M., 2016. The promise of ultra-short-term (UST) heart rate variability measurements. Biofeedback 44 (4), 229–233. https://doi.org/10.5298/1081-5937-44.3.09.
Shaffer, F., Meehan, Z.M., Zerr, C.L., 2020. A critical review of ultra-short-term heart rate variability norms research. Front. Neurosci. 14, 594880 https://doi.org/10.3389/fnins.2020.594880.
Shcheslavskaya, O.V., Burg, M.M., McKinley, P.S., Schwartz, J.E., Gerin, W., Ryff, C.D., Weinstein, M., Seeman, T.E., Sloan, R.P., 2010. Heart rate recovery after cognitive challenge is preserved with age. Psychosom. Med. 72 (2), 128–133. https://doi.org/10.1097/PSY.0b013e3181c94ca0.
Shen, W., Yuan, Y., Tang, C., Shi, C., Liu, C., Luo, J., Zhang, X., 2017. In search of somatic precursors of spontaneous insight. J. Psychophysiol. 32 (3), 97–105. https://doi.org/10.1027/0269-8803/a000188.
Silverthorn, D.U., 2004. Human Physiology: An Integrated Approach with Interactive Physiology, Third edition. Pearson Education, Inc., San Francisco.
So, T.Y., Li, M.Y.E., Lau, H., 2021. Between-subject correlation of heart rate variability predicts movie preferences. PloS One 16 (2), e0247625. https://doi.org/10.1371/journal.pone.0247625.
Stone, J.D., Ulman, H.K., Tran, K., Thompson, A.G., Halter, M.D., Ramadan, J.H., Hagen, J.A., 2021. Assessing the accuracy of popular commercial technologies that measure resting heart rate and heart rate variability. Front. Sports Active Living 3, 37. https://doi.org/10.3389/fspor.2021.585870.
Tarvainen, M.P., Ranta-Aho, P.O., Karjalainen, P.A., 2002. An advanced detrending method with application to HRV analysis. IEEE Trans. Biomed. Eng. 49 (2), 172–175. https://doi.org/10.1109/10.979357.
Tarvainen, M.P., Niskanen, J.P., Lipponen, J.A., Ranta-Aho, P.O., Karjalainen, P.A., 2014. Kubios HRV–heart rate variability analysis software. Comput. Methods Prog. Biomed. 113 (1), 210–220. https://doi.org/10.1016/j.cmpb.2013.07.024.
Tarvainen, M.P., Lipponen, J., Niskanen, J.P., Ranta-aho, P.O., 2020. Kubios HRV User Guide, Kubios Oy.
Task Force of the European Society of Cardiology the North American Society of Pacing Electrophysiology, 1996. Heart rate variability: Standards of measurment, physiological interpretation, and clinical use. Circulation 93 (5), 1043–1065. https://doi.org/10.1161/01.CIR.93.5.1043.
Torchiano, M., 2020. effsize: Efficient Effect Size Computation (R Package Version 0.8.1). https://doi.org/10.5281/zenodo.1480624.
Umetani, K., Singer, D.H., McCraty, R., Atkinson, M., 1998. Twenty-four hour time domain heart rate variability and heart rate: relations to age and gender over nine decades. J. Am. Coll. Cardiol. 31 (3), 593–601. https://doi.org/10.1016/s0735-1097(97)00554-8.
van Lier, H.G., Pieterse, M.E., Garde, A., Postel, M.G., de Haan, H.A., Vollenbroek-Hutten, M.M., Noordzij, M.L., 2020. A standardized validity assessment protocol for physiological signals from wearable technology: methodological underpinnings and an application to the E4 biosensor. Behav. Res. Methods 52 (2), 607–629. https://doi.org/10.3758/s13428-019-01263-9.
Yuda, E., Shibata, M., Ogata, Y., Ueda, N., Yambe, T., Yoshizawa, M., Hayano, J., 2020. Pulse rate variability: a new biomarker, not a surrogate for heart rate variability. J. Physiol. Anthropol. 39 (1), 21. https://doi.org/10.1186/s40101-020-00233-x.